23.2.2 Give a rule for continuing the sequence

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{3}{1}, \frac{2}{2}, \frac{1}{3}, \frac{4}{1}, \frac{3}{2}, \cdots$$

so as to include all positive rationals.

23.2.3 How can one then conclude that the set of all rationals is countable?

23.2.4 The words on a fixed finite alphabet can be enumerated by listing first the one-letter words, then the two-letter words, and so on. Use this observation to show that the set of polynomial equations with integer coefficients is countable and hence that the set of algebraic numbers is countable.

Cantor used the latter result to prove the existence of transcendental numbers. Namely, let $\{x_m\}$ be the sequence of algebraic numbers; we know that these are not all the real numbers, so any other real number is transcendental.

23.3 Measure

The reason for investigating sets of discontinuities in the theory of Fourier series was the discovery of Fourier (1822) that these series depend on integrals. Assuming that

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} (a_n \cos n\pi x + b_n \sin n\pi x),$$

Fourier derived the formulas

$$a_n = \int_{-1}^{1} f(x) \cos n\pi x \, dx, \quad b_n = \int_{-1}^{1} f(x) \sin n\pi x \, dx.$$

Thus the existence of the series depends on the existence of the integrals for a_n and b_n, and this in turn depends on how discontinuous f is. It was known (though not rigorously proved) that every continuous function had an integral, so the next question was how the integral should, or could, be defined for discontinuous functions. The first precise answer was the Riemann (1854a) integral concept, familiar to all calculus students, and based on approximating the integrand by step functions. Any function with a finite number of discontinuities has a Riemann integral, and indeed so have certain functions with infinitely many discontinuities, but not all. The classic function for which the Riemann integral does not exist is the function of Dirichlet (1829):

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

Eventually a more general integral, the Lebesgue integral, was introduced to cope with such functions, but not until the focus of attention had shifted from the problem of integration to the more fundamental problem of *measure*. Measure generalizes the concept of length (on the line $\mathbb{R}$), area (in the plane $\mathbb{R}^2$), and so on, to quite general point sets. Since an integral can be viewed as the area under a graph, its dependence on the concept of measure is clear, though it was not immediately realized that the measure of sets on the line had to be clarified first.

The need for clarification arose from the discovery of Harnack (1885) that any countable subset $\{x_0, x_1, x_2, \ldots\}$ of $\mathbb{R}$ could be covered by a collection of intervals of arbitrarily small total length (cover x_0 by an interval of length $\varepsilon/2$, x_1 by an interval of length $\varepsilon/4$, x_2 by an interval of length $\varepsilon/8$, ..., so that the total length of intervals used is $\leq \varepsilon$). This seemed to show that countable sets were "small" (of *measure zero*, as we now say), but mathematicians were reluctant to say this of dense countable sets, like the rationals. The first response was to define measure analogously to the Riemann integral, using finite unions of intervals to approximate subsets of $\mathbb{R}$ [Jordan (1892)]. Under this definition, "sparse" countable sets like $\{0, 1/2, 3/4, 7/8, \ldots\}$ did have measure zero, but dense sets like the rationals were not measurable at all.

The first to take the hint from Harnack's result that countable unions of intervals should be used to measure subsets of $\mathbb{R}$ was Borel (1898). He defined the measure of any countable union of intervals to be its total length, and he extended measurability to more and more complicated sets by *complementation* and *countable disjoint unions*. That is, if a set S contained in an interval I has measure $\mu(S)$, then

$$\mu(I - S) = \mu(I) - \mu(S),$$

and if S is a disjoint union of sets S_n with measures $\mu(S_n)$, then

$$\mu(S) = \sum_{n=1}^{\infty} \mu(S_n).$$

The sets that can be formed from intervals by complementation and countable unions are now called *Borel sets*. Borel's idea was pushed to its logical conclusion by Lebesgue (1902), who assigned measure zero to any subset of a Borel set of measure zero. Since not all such sets are Borel, this extended measurability to a larger class of sets: those that differ from Borel's

by sets of measure zero. Whether the measurable sets are *all* subsets of $\mathbb{R}$ is an interesting question to which we shall return shortly.

The distinctive property of Borel–Lebesgue measure is *countable additivity*: if $S_0, S_1, S_2, \ldots$ are disjoint measurable sets, then

$$\mu\left(S_0 \cup S_1 \cup S_2 \cup \cdots\right) = \mu\left(S_0\right) + \mu\left(S_1\right) + \mu\left(S_2\right) + \cdots.$$

Lebesgue showed that this gives a concept of integral that is better behaved with respect to limits than the Riemann integral. For example, one has the *monotone convergence property*: if $f_0, f_1, f_2, \ldots$ is an increasing sequence of positive integrable functions, and $f_n \to f$ as $n \to \infty$, then $\int f_n \, dx \to \int f \, dx$ for the Lebesgue integral, whereas this is not generally true for the Riemann integral (see Exercise 23.3.1).

Another motivation for countable additivity Borel pointed out was the theory of probability. If an "event" E is formalized as a set S of points ("favorable outcomes"), then the probability of E can be defined as the measure of S. Some quite natural events turn out to be countable unions, hence it is necessary for probability measure to be countably additive. In informal probability theory, countable additivity was assumed as far back as 1690, when Jakob Bernoulli answered the following question he had posed in 1685:

> A and B play with a die, the one that throws an ace first being declared the winner. A throws once, then B throws once also. A then throws twice, and B does the same, and so on, until one wins. What is the ratio of their chances of success?

To solve this problem, Jakob Bernoulli (1690) decomposed the event of a win for A (or B) into the subevents of a win at A's (B's) first, second, third, ..., turn and summed the probabilities of these countably many subevents. Formal probability theory, which was created by Kolmogorov (1933), bases all such arguments on the theory of countably additive measures.

It could be said that set theory paved the way for measure theory by showing the uncountability of $\mathbb{R}$, thus enabling countable subsets of $\mathbb{R}$ to be regarded as "small." On the other hand, measure theory itself shows the uncountability of $\mathbb{R}$ (look again at Harnack's result), and in fact measure theory's assessment of the "smallness" of countable sets greatly influenced the later development of set theory.

"Measure theoretically desirable" axioms, such as the measurability of all subsets of $\mathbb{R}$, turned out to conflict with "set theoretically desirable" axioms such as the continuum hypothesis, and efforts to resolve the conflict brought more fundamental questions about sets to light. These questions do not reduce to clear-cut alternatives—the way geometric questions reduce to alternative parallel axioms, for example—but they do seem to gravitate toward the so-called *choice* and *large cardinal* axioms, discussed in the next section.

EXERCISES

23.3.1 Show that a function f_n that is zero at all but n points has Riemann integral zero over any interval and that the non-Riemann integrable function of Dirichlet is a limit as $n \to \infty$ of such functions f_n.

The complexity of Borel sets may be roughly measured by the number of countable unions and complements needed to define them. Here are a few of the simpler ones.

23.3.2 Show that a single point is the complement of a countable union of intervals and hence that any countable set is a Borel set.

23.3.3 Deduce that the set of irrational numbers is a Borel set.

23.3.4 What is the measure of the set of irrationals between 0 and 1?

23.4 Axiom of Choice and Large Cardinals

The usual axiom of choice states that for any set S (of sets) there is a *choice function* f such that $f(x) \in x$ for each $x \in S$. (Thus f "chooses" an element from each set x in S.) The axiom seems so plausible that early set theorists used it almost unconsciously, and it first attracted attention in Zermelo's (1904) proof that any set S could be *well ordered* (that is, put in one-to-one correspondence with an ordinal). This looked like progress toward the continuum hypothesis. But Zermelo's proof gave no more than the existence of a well-ordering of S, given a choice function for the set of subsets of S. There was still no sign of an explicit well-ordering of $\mathbb{R}$. And of course if one doubted the existence of a well-ordering of $\mathbb{R}$, this threw doubt on the axiom of choice. Further doubts were raised when the axiom of choice was found to have incredible consequences in measure theory.

The first of these, discovered by Vitali (1905), was that the circle can be decomposed into countably many disjoint congruent sets. Since congruent

sets have the same Lebesgue measure, it easily follows that the sets in question are not Lebesgue measurable (by countable additivity; see Exercises 23.4.2–23.4.4).

Even more paradoxical decompositions were given by Hausdorff (1914) (for the sphere) and Banach and Tarski (1924) (for the ball). The Banach–Tarski theorem states that the unit ball can be decomposed into finitely many sets that, when rigidly moved in space, form *two* unit balls! This shows that not all subsets of the ball are measurable, even if one asks only for finite, rather than countable, additivity. For an excellent discussion of the paradoxical decompositions and their connections with other parts of mathematics, see Wagon (1985).

The measure-theoretic consequences of the paradoxical decompositions follow from the geometrically natural assumption that congruent sets have the same measure. If one drops this assumption and asks only for countable additivity and nontriviality (that is, not all subsets have measure zero), then the conflict with the axiom of choice seems to disappear. No contradiction has yet been derived from these assumptions, but Ulam (1930) showed that any set possessing such a measure must be extraordinarily large—large enough, in fact, to be a model for set theory itself, and in particular larger than the cardinals $\aleph_1, \aleph_2, \ldots, \aleph_\omega, \ldots$. Thus if $\mathbb{R}$ has a nontrivial countably additive measure, then $\mathbb{R}$ must be far larger than $\aleph_1$, and we still have a conflict with the continuum hypothesis. (For more on the "largeness" of models, see Section 23.8.)

An even more desirable axiom than measurability would be Lebesgue measurability of all subsets of $\mathbb{R}$. This conflicts with the axiom of choice, by Vitali's theorem, but it was nevertheless shown to be consistent with set theory by Solovay (1970), assuming the existence of a large cardinal. Shelah (1984) showed that the large cardinal assumption is necessary.

Thus measurability of all subsets of $\mathbb{R}$ is intimately connected with the existence of sets large enough to model the whole of set theory. This mind-boggling concept seems to be the answer to many fundamental questions. We shall find ourselves drawn to it again in the next sections when we explore the influence of set theory on logic. Meanwhile, for those who would like a more detailed account of the development of set theory, and the contentious axioms in particular, we refer to van Dalen and Monna (1972). For recent developments in the theory of large cardinals, which some believe will throw new light on the continuum hypothesis, see Kanamori (1994) and Woodin (1999).

EXERCISES

The axiom of choice turns up even in elementary analysis, when one attempts to formalize the idea of a continuous function. A natural definition in terms of infinite sequences is equivalent to the standard ε-δ definition only if we assume the axiom of choice.

Call f *sequentially continuous* at $x = a$ if, for any sequence $\{a_n\}$ such that $a_n \to a$, we have $f(a_n) \to f(a)$.

23.4.1 Show, assuming the axiom of choice, that if f is not continuous at a then f is not sequentially continuous at a. [It is a consequence of Cohen (1963) that this result *cannot* be proved without the axiom of choice.]

Vitali's decomposition of the circle is created as follows. For each θ between 0 and 2π let $S(\theta)$ be the set of points on the unit circle whose angle differs from θ by a rational multiple of 2π. Thus $S(\theta) = S(\phi)$ if $\theta - \phi = 2\pi \times$ a rational, and $S(\theta) \cap S(\phi) = \emptyset$ otherwise.

23.4.2 Let S be a set (existing by virtue of the axiom of choice) that contains exactly one element from each distinct $S(\theta)$ and let

$$S + 2\pi r = \{\theta + 2\pi r : \theta \in S\} \quad \text{for each rational } r.$$

(Thus $S + 2\pi r$ is S rotated through the rational multiple $2\pi r$ of 2π.) Show that any two of the sets $S + 2\pi r$ are either identical or disjoint.

23.4.3 Show that the circle is a countable union of sets $S + 2\pi r$.

23.4.4 Show that both assumptions $\mu(S) = 0$ and $\mu(S) > 0$ lead to contradictions, and hence conclude that S is nonmeasurable.

23.5 The Diagonal Argument

The uncountability of $\mathbb{R}$ was shown again in a strikingly simple way by Cantor (1891). His argument applies most directly to the set $2^{\mathbb{N}}$ of all subsets of $\mathbb{N}$, but there are variants that work similarly on the set $\mathbb{N}^{\mathbb{N}}$ of integer functions and on $\mathbb{R}$ (which can be identified with a set of integer functions in various ways). To show that there are uncountably many subsets of $\mathbb{N}$ one shows that any countable collection $S_0, S_1, S_2, \ldots$ of sets $S_n \subseteq \mathbb{N}$ is incomplete, by constructing a new set S, different from each S_n. S is the so-called *diagonal set* $\{n : n \notin S_n\}$, which obviously differs from S_n with respect to the number n. Q.E.D.

The "diagonal" nature of S can be seen by visualizing a table of 0's and 1's in which

$$m\text{th entry in } n\text{th row} = \begin{cases} 0 & \text{if } m \notin S_n, \\ 1 & \text{if } m \in S_n. \end{cases}$$

In other words, the nth row consists of the values of the characteristic function of S_n. The characteristic function of S is simply the diagonal of the table, with all values reversed. A sequence $x_0, x_1, x_2, \ldots$ of real numbers can be diagonalized similarly by forming the table whose nth row consists of the decimal digits of x_n. A suitable way to "reverse" the digits on the diagonal is to change any 1 to a 2 and any other digit to a 1. (The resulting sequence of 1's and 2's, after a decimal point, then defines a real number x whose decimal expansion is unique. Hence x is not just different from each x_n in its decimal expansion but is definitely a different number.)

More generally, for any table of rows of integers, that is, any sequence of integer functions f_n, one can construct an integer function f unequal to each f_n by changing the values along the diagonal of the table. The diagonal argument was in fact first given in this context, by du Bois-Reymond (1875), in order to construct an f with a greater rate of growth than all functions in a sequence $f_0, f_1, f_2, \ldots$ (Exercise 23.5.1). With hindsight, one can even see a diagonal construction in Cantor's first (1874) argument for the uncountability of $\mathbb{R}$ (Exercise 23.5.2).

The diagonal argument is important in set theory because it readily generalizes to show that any set has more subsets than elements (Exercise 23.5.3), and hence that there is no largest set. What was not noticed at first is that the diagonal argument also has consequences at a more concrete level. This is because the diagonal of a table is *computable* if the table as a whole is computable. Hence the argument does not merely show how to add a new function f to a list $f_0, f_1, f_2, \ldots$—it shows how to add a new computable function to a computable list. In other words, it is *impossible to compute a list of all computable functions.* And of course the same goes for lists of computable real numbers. This remarkable result went unnoticed in the early days of the diagonal argument because computability was not then regarded as an interesting concept, or indeed as a mathematical concept at all. The controversies over the axiom of choice, however, helped to sharpen awareness of the difference between constructive and nonconstructive functions. In the 1920s logicians began to investigate the concept of computability more seriously, and by a "kind of miracle," as Gödel (1946) later expressed it, computability turned out to be a mathematically precise notion.

EXERCISES

The diagonal construction is quite a natural way to construct a function or real number "larger" than the members of a given countable set.