CHAPTER 1

Probability

SECTION 1. BOREL'S NORMAL NUMBER THEOREM

Although sufficient for the development of many interesting topics in mathematical probability, the theory of discrete probability spaces[†] does not go far enough for the rigorous treatment of problems of two kinds: those involving an infinitely repeated operation, as an infinite sequence of tosses of a coin, and those involving an infinitely fine operation, as the random drawing of a point from a segment. A mathematically complete development of probability, based on the theory of measure, puts these two classes of problem on the same footing, and as an introduction to measure-theoretic probability it is the purpose of the present section to show by example why this should be so.

The Unit Interval

The project is to construct simultaneously a model for the random drawing of a point from a segment and a model for an infinite sequence of tosses of a coin. The notions of independence and expected value, familiar in the discrete theory, will have analogues here, and some of the terminology of the discrete theory will be used in an informal way to motivate the development. The formal mathematics, however, which involves only such notions as the length of an interval and the Riemann integral of a step function, will be entirely rigorous. All the ideas will reappear later in more general form.

Let Ω denote the unit interval $(0, 1]$; to be definite, take intervals open on the left and closed on the right. Let ω denote the generic point of Ω. Denote the length of an interval $I = (a, b]$ by $|I|$:

$$|I| = |(a, b]| = b - a. \tag{1.1}$$

[†] For the discrete theory, presupposed here, see for example the first half of Volume I of FELLER. (Names in capital letters refer to the bibliography on p. 581.)

If

$$A = \bigcup_{i=1}^{n} I_i = \bigcup_{i=1}^{n} (a_i, b_i], \tag{1.2}$$

where the intervals $I_i = (a_i, b_i]$ are disjoint [A3]† and are contained in Ω, assign to A the probability

$$P(A) = \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} (b_i - a_i). \tag{1.3}$$

It is important to understand that in this section $P(A)$ is defined only if A is a finite disjoint union of subintervals of $(0, 1]$—never for sets A of any other kind.

If A and B are two such finite disjoint unions of intervals, and if A and B are disjoint, then $A \cup B$ is a finite disjoint union of intervals and

$$P(A \cup B) = P(A) + P(B). \tag{1.4}$$

This relation, which is certainly obvious intuitively, is a consequence of the additivity of the Riemann integral:

$$\int_0^1 (f(\omega) + g(\omega))\, d\omega = \int_0^1 f(\omega)\, d\omega + \int_0^1 g(\omega)\, d\omega. \tag{1.5}$$

If $f(\omega)$ is a step function taking value c_j in the interval $(x_{j-1}, x_j]$, where $0 = x_0 < x_1 < \cdots < x_k = 1$, then its integral in the sense of Riemann has the value

$$\int_0^1 f(\omega)\, d\omega = \sum_{j=1}^{k} c_j (x_j - x_{j-1}). \tag{1.6}$$

If $f = I_A$ and $g = I_B$ are the indicators [A5] of A and B, then (1.4) follows from (1.5) and (1.6), provided A and B are disjoint. This also shows that the definition (1.3) is unambiguous—note that A will have many representations of the form (1.2) because $(a, b] \cup (b, c] = (a, c]$. Later these facts will be derived anew from the general theory of Lebesgue integration.‡

According to the usual models, if a radioactive substance has emitted a single α-particle during a unit interval of time, or if a single telephone call has arrived at an exchange during a unit interval of time, then the instant at which the emission or the arrival occurred is random in the sense that it lies in (1.2) with probability (1.3). Thus (1.3) is the starting place for the description of a point drawn at random from the unit interval: Ω is regarded as a sample space, and the set (1.2) is identified with the event that the random point lies in it.

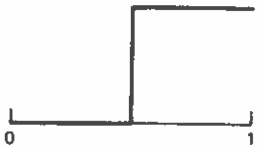
The definition (1.3) is also the starting point for a mathematical representation of an infinite sequence of tosses of a coin. With each ω associate its nonterminating dyadic expansion

$$\omega = \sum_{n=1}^{\infty} \frac{d_n(\omega)}{2^n} = .d_1(\omega) d_2(\omega) \ldots, \tag{1.7}$$

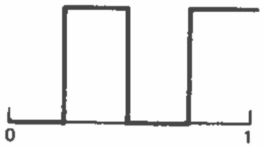
each $d_n(\omega)$ being 0 or 1 [A31]. Thus

$$(d_1(\omega), d_2(\omega), \ldots) \tag{1.8}$$

is the sequence of binary digits in the expansion of ω. For definiteness, a point such as $\frac{1}{2} = .1000\ldots = .0111\ldots$, which has two expansions, takes the nonterminating one; 1 takes the expansion $.111\ldots$.



Graph of $d_1(\omega)$



Graph of $d_2(\omega)$

Imagine now a coin with faces labeled 1 and 0 instead of the usual heads and tails. If ω is drawn at random, then (1.8) behaves as if it resulted from an infinite sequence of tosses of a coin. To see this, consider first the set of ω for which $d_i(\omega) = u_i$ for $i = 1, \ldots, n$, where $u_1, \ldots, u_n$ is a sequence of 0's and 1's. Such an ω satisfies

$$\sum_{i=1}^{n} \frac{u_i}{2^i} < \omega \le \sum_{i=1}^{n} \frac{u_i}{2^i} + \sum_{i=n+1}^{\infty} \frac{1}{2^i},$$

where the extreme values of ω correspond to the case $d_i(\omega) = 0$ for $i > n$ and the case $d_i(\omega) = 1$ for $i > n$. The second case can be achieved, but since the binary expansions represented by the $d_i(\omega)$ are nonterminating—do not end in 0's—the first cannot, and ω must actually exceed $\sum_{i=1}^{n} u_i/2^i$. Thus

$$[\omega: d_i(\omega) = u_i, i = 1, \ldots, n] = \left(\sum_{i=1}^{n} \frac{u_i}{2^i}, \sum_{i=1}^{n} \frac{u_i}{2^i} + \frac{1}{2^n} \right]. \tag{1.9}$$

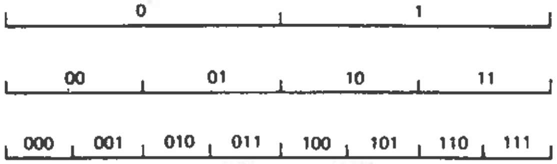
†A notation [An] refers to paragraph n of the appendix beginning on p. 536; this is a collection of mathematical definitions and facts required in the text.

‡Passages in small type concern side issues and technical matters, but their contents are sometimes required later.

The interval here is open on the left and closed on the right precisely because the expansion (1.7) is the nonterminating one. In the model for coin tossing the set (1.9) represents the event that the first n tosses give the outcomes $u_1, \ldots, u_n$ in sequence. By (1.3) and (1.9),

$$(1.10) \qquad P[\omega: d_i(\omega) = u_i,\ i = 1, \ldots, n] = \frac{1}{2^n},$$

which is what probabilistic intuition requires.



Decompositions by dyadic intervals

The intervals (1.9) are called *dyadic* intervals, the endpoints being adjacent dyadic rationals $k/2^n$ and $(k+1)/2^n$ with the same denominator, and n is the *rank* or *order* of the interval. For each n the 2^n dyadic intervals of rank n decompose or partition the unit interval. In the passage from the partition for n to that for $n+1$, each interval (1.9) is split into two parts of equal length, a left half on which $d_{n+1}(\omega)$ is 0 and a right half on which $d_{n+1}(\omega)$ is 1. For $u = 0$ and for $u = 1$, the set $[\omega: d_{n+1}(\omega) = u]$ is thus a disjoint union of 2^n intervals of length $1/2^{n+1}$ and hence has probability $\frac{1}{2}$: $P[\omega: d_n(\omega) = u] = \frac{1}{2}$ for all n.

Note that $d_i(\omega)$ is constant over each dyadic interval of rank i and that for $n > i$ each dyadic interval of rank n is entirely contained in a single dyadic interval of rank i. Therefore, $d_i(\omega)$ is constant over each dyadic interval of rank n if $i \le n$.

The probabilities of various familiar events can be written down immediately. The sum $\sum_{i=1}^n d_i(\omega)$ is the number of 1's among $d_1(\omega), \ldots, d_n(\omega)$, to be thought of as the number of heads in n tosses of a fair coin. The usual binomial formula is

$$(1.11) \qquad P\left[\omega: \sum_{i=1}^{n} d_i(\omega) = k\right] = \binom{n}{k}\frac{1}{2^n}, \qquad 0 \le k \le n.$$

This follows from the definitions: The set on the left in (1.11) is the union of those intervals (1.9) corresponding to sequences $u_1, \ldots, u_n$ containing k 1's and $n - k$ 0's; each such interval has length $1/2^n$ by (1.10) and there are $\binom{n}{k}$ of them, and so (1.11) follows from (1.3).

The functions $d_n(\omega)$ can be looked at in two ways. Fixing n and letting ω vary gives a real function $d_n = d_n(\cdot)$ on the unit interval. Fixing ω and letting n vary gives the sequence (1.8) of 0's and 1's. The probabilities (1.10) and (1.11) involve only finitely many of the components $d_i(\omega)$. The interest here, however, will center mainly on properties of the entire sequence (1.8). It will be seen that the mathematical properties of this sequence mirror the properties to be expected of a coin-tossing process that continues forever.

As the expansion (1.7) is the nonterminating one, there is the defect that for no ω is (1.8) the sequence $(1, 0, 0, 0, \ldots)$, for example. It seems clear that the chance should be 0 for the coin to turn up heads on the first toss and tails forever after, so that the absence of $(1, 0, 0, 0, \ldots)$—or of any other single sequence—should not matter. See on this point the additional remarks immediately preceding Theorem 1.2.

The Weak Law of Large Numbers

In studying the connection with coin tossing it is instructive to begin with a result that can, in fact, be treated within the framework of discrete probability, namely, the *weak law of large numbers*:

Theorem 1.1. *For each* ϵ,[†]

$$(1.12) \qquad \lim_{n\to\infty} P\left[\omega: \left|\frac{1}{n}\sum_{i=1}^{n} d_i(\omega) - \frac{1}{2}\right| \ge \epsilon\right] = 0.$$

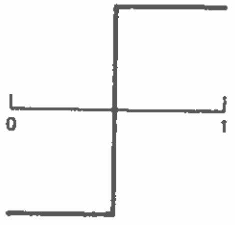
Interpreted probabilistically, (1.12) says that if n is large, then there is small probability that the fraction or relative frequency of heads in n tosses will deviate much from $\frac{1}{2}$, an idea lying at the base of the frequency conception of probability. As a statement about the structure of the real numbers, (1.12) is also interesting arithmetically.

Since $d_i(\omega)$ is constant over each dyadic interval of rank n if $i \le n$, the sum $\sum_{i=1}^n d_i(\omega)$ is also constant over each dyadic interval of rank n. The set in (1.12) is therefore the union of certain of the intervals (1.9), and so its probability is well defined by (1.3).

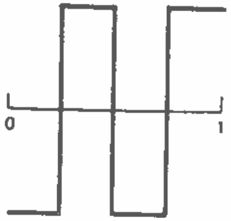
With the Riemann integral in the role of expected value, the usual application of Chevyshev's inequality will lead to a proof of (1.12). The argument becomes simpler if the $d_n(\omega)$ are replaced by the *Rademacher functions*,

$$(1.13) \qquad r_n(\omega) = 2d_n(\omega) - 1 = \begin{cases} +1 & \text{if } d_n(\omega) = 1, \\ -1 & \text{if } d_n(\omega) = 0. \end{cases}$$

[†] The standard ϵ and δ of analysis will always be understood to be positive.



Graph of $r_1(\omega)$



Graph of $r_2(\omega)$

Consider the partial sums

(1.14) $$s_n(\omega) = \sum_{i=1}^{n} r_i(\omega).$$

Since $\Sigma_{i=1}^n d_i(\omega) = (s_n(\omega)+n)/2$, (1.12) with $\epsilon/2$ in place of ϵ is the same thing as

(1.15) $$\lim_{n\to\infty} P\left[\omega: \left|\frac{1}{n}s_n(\omega)\right| \ge \epsilon\right] = 0.$$

This is the form in which the theorem will be proved.

The Rademacher functions have themselves a direct probabilistic meaning. If a coin is tossed successively, and if a particle starting from the origin performs a random walk on the real line by successively moving one unit in the positive or negative direction according as the coin falls heads or tails, then $r_i(\omega)$ represents the distance it moves on the ith step and $s_n(\omega)$ represents its position after n steps. There is also the gambling interpretation: If a gambler bets one dollar, say, on each toss of the coin, $r_i(\omega)$ represents his gain or loss on the ith play and $s_n(\omega)$ represents his gain or loss in n plays.

Each dyadic interval of rank $i-1$ splits into two dyadic intervals of rank i; $r_i(\omega)$ has value -1 on one of these and value $+1$ on the other. Thus $r_i(\omega)$ is -1 on a set of intervals of total length $\frac{1}{2}$ and $+1$ on a set of total length $\frac{1}{2}$. Hence $\int_0^1 r_i(\omega)\,d\omega = 0$ by (1.6), and

(1.16) $$\int_0^1 s_n(\omega)\,d\omega = 0$$

by (1.5). If the integral is viewed as an expected value, then (1.16) says that the mean position after n steps of a random walk is 0.

Suppose that $i<j$. On a dyadic interval of rank $j-1$, $r_i(\omega)$ is constant and $r_j(\omega)$ has value -1 on the left half and $+1$ on the right. The product $r_i(\omega)r_j(\omega)$ therefore integrates to 0 over each of the dyadic intervals of rank $j-1$, and so

(1.17) $$\int_0^1 r_i(\omega)r_j(\omega)\,d\omega = 0, \qquad i \ne j.$$

This corresponds to the fact that independent random variables are uncorrelated. Since $r_i^2(\omega)=1$, expanding the square of the sum (1.14) shows that

(1.18) $$\int_0^1 s_n^2(\omega)\,d\omega = n.$$

This corresponds to the fact that the variances of independent random variables add. Of course (1.16), (1.17), and (1.18) stand on their own, in no way depend on any probabilistic interpretation.

Applying Chebyshev's inequality in a formal way to the probability in (1.15) now leads to

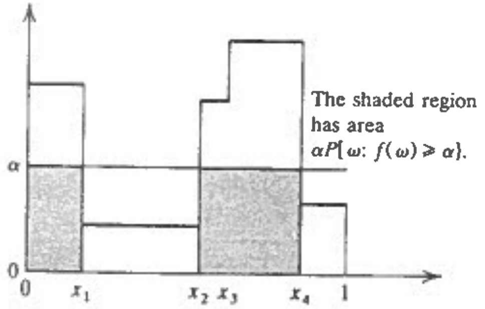
(1.19) $$P[\omega: |s_n(\omega)| \ge n\epsilon] \le \frac{1}{n^2\epsilon^2}\int_0^1 s_n^2(\omega)\,d\omega = \frac{1}{n\epsilon^2}.$$

The following lemma justifies the inequality.

Let f be a step function as in (1.6): $f(\omega) = c_j$ for $\omega \in (x_{j-1}, x_j]$, where $0 = x_0 < \cdots < x_k = 1$.

Lemma. *If f is a nonnegative step function, then $[\omega: f(\omega) \ge \alpha]$ is for $\alpha > 0$ a finite union of intervals and*

(1.20) $$P[\omega: f(\omega) \ge \alpha] \le \frac{1}{\alpha}\int_0^1 f(\omega)\,d\omega.$$



The shaded region has area $\alpha P[\omega: f(\omega) \geqslant \alpha]$.

Proof. The set in question is the union of the intervals $(x_{j-1}, x_j]$ for which $c_j \ge \alpha$. If Σ' denotes summation over those j satisfying $c_j \ge \alpha$, then $P[\omega: f(\omega) \ge \alpha] = \Sigma'(x_j - x_{j-1})$ by the definition (1.3). On the other hand, since the c_j are all nonnegative by hypothesis, (1.6) gives

$$\int_0^1 f(\omega)\,d\omega = \sum_{j=1}^{k} c_j(x_j - x_{j-1}) \ge \sum{}' c_j(x_j - x_{j-1})$$
$$\ge \sum{}' \alpha(x_j - x_{j-1}).$$

Hence (1.20). $\blacksquare$

Taking $\alpha = n^2\epsilon^2$ and $f(\omega) = s_n^2(\omega)$ in (1.20) gives (1.19). Clearly, (1.19) implies (1.15), and as already observed, this in turn implies (1.12).

The Strong Law of Large Numbers

It is possible with a minimum of technical apparatus to prove a stronger result that cannot even be formulated in the discrete theory of probability. Consider the set

$$N = \left[\omega:\ \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} d_i(\omega) = \frac{1}{2}\right] \tag{1.21}$$

consisting of those ω for which the asymptotic relative frequency[*] of 1 in the sequence (1.8) is $\frac{1}{2}$. The points in (1.21) are called *normal numbers*. The idea is to show that a real number ω drawn at random from the unit interval is "practically certain" to be normal, or that there is "practical certainty" that 1 occurs in the sequence (1.8) of tosses with asymptotic relative frequency $\frac{1}{2}$. It is impossible at this stage to prove that $P(N) = 1$, because N is not a finite union of intervals and so has been assigned no probability. But the notion of "practical certainty" can be formalized in the following way.

Define a subset A of Ω to be *negligible*[†] if for each positive ϵ there exists a finite or countable[‡] collection $I_1, I_2, \ldots$ of intervals (they may overlap) satisfying

$$A \subset \bigcup_k I_k \tag{1.22}$$

and

$$\sum_k |I_k| < \epsilon. \tag{1.23}$$

A negligible set is one that can be covered by intervals the total sum of whose lengths can be made arbitrarily small. If $P(A)$ is assigned to such an

A in any reasonable way, then for the I_k of (1.22) and (1.23) it ought to be true that $P(A) \le \sum_k P(I_k) = \sum_k |I_k| < \epsilon$, and hence $P(A)$ ought to be 0. Even without any assignment of probability at all, the definition of negligibility can serve as it stands as an explication of "practical impossibility" and "practical certainty": Regard it as practically impossible that the random ω will lie in A if A is negligible, and regard it as practically certain that ω will lie in A if its complement A^c [A1] is negligible.

Although the fact plays no role in the next proof, for an understanding of negligibility observe first that *a finite or countable union of negligible sets is negligible*. Indeed, suppose that $A_1, A_2, \ldots$ are negligible. Given ϵ, for each n choose intervals $I_{n1}, I_{n2}, \ldots$ such that $A_n \subset \bigcup_k I_{nk}$ and $\sum_k |I_{nk}| < \epsilon/2^n$. All the intervals I_{nk} taken together form a countable collection covering $\bigcup_n A_n$, and their lengths add to $\sum_n \sum_k |I_{nk}| < \sum_n \epsilon/2^n = \epsilon$. Therefore, $\bigcup_n A_n$ is negligible.

A set consisting of a single point is clearly negligible, and so every countable set is also negligible. The rationals for example form a negligible set. In the coin-tossing model, a single point of the unit interval has the role of a single sequence of 0's and 1's, or of a single sequence of heads and tails. It corresponds with intuition that it should be "practically impossible" to toss a coin infinitely often and realize any one particular infinite sequence set down in advance. It is for this reason not a real shortcoming of the model that for no ω is (1.8) the sequence $(1, 0, 0, 0, \ldots)$. In fact, since a countable set is negligible, it is not a shortcoming that (1.8) is never one of the countably many sequences that end in 0's.

Theorem 1.2. *The set of normal numbers has negligible complement.*

This is *Borel's normal number theorem*,[†] a special case of the *strong law of large numbers*. Like Theorem 1.1, it is of arithmetic as well as probabilistic interest.

The set N^c is not countable: Consider a point ω for which $(d_1(\omega), d_2(\omega), \ldots) = (1, 1, u_3, 1, 1, u_6, \ldots)$—that is, a point for which $d_i(\omega) = 1$ unless i is a multiple of 3. Since $n^{-1}\sum_{i=1}^{n} d_i(\omega) \ge \frac{2}{3}$, such a point cannot be normal. But there are uncountably many such points, one for each infinite sequence $(u_3, u_6, \ldots)$ of 0's and 1's. Thus one cannot prove N^c negligible by proving it countable, and a deeper argument is required.

PROOF OF THEOREM 1.2. Clearly (1.21) and

$$N = \left[\omega:\ \lim_{n\to\infty} \frac{1}{n} s_n(\omega) = 0\right] \tag{1.24}$$

[*] The *frequency* of 1 (the number of occurrences of it) among $d_1(\omega), \ldots, d_n(\omega)$ is $\sum_{i=1}^{n} d_i(\omega)$, the *relative* frequency is $n^{-1}\sum_{i=1}^{n} d_i(\omega)$, and the *asymptotic* relative frequency is the limit in (1.21).

[†] The term *negligible* is introduced for the purposes of this section only. The negligible sets will reappear later as the sets of Lebesgue measure 0.

[‡] *Countably infinite* is unambiguous. *Countable* will mean finite or countably infinite, although it will sometimes for emphasis be expanded as here to *finite or countable*.

[†] Émile Borel: Sur les probabilités dénombrables et leurs applications arithmétiques, *Circ. Mat. d. Palermo*, **29** (1909), 247–271. See DUDLEY for excellent historical notes on analysis and probability.

define the same set (see (1.14)). To prove N^c negligible requires constructing coverings that satisfy (1.22) and (1.23) for $A = N^c$. The construction makes use of the inequality

$$P[\omega: |s_n(\omega)| \ge n\epsilon] \le \frac{1}{n^4\epsilon^4}\int_0^1 s_n^4(\omega)\,d\omega. \tag{1.25}$$

This follows by the same argument that leads to the inequality in (1.19)—it is only necessary to take $f(\omega) = s_n^4(\omega)$ and $\alpha = n^4\epsilon^4$ in (1.20). As the integral in (1.25) will be shown to have order n^2, the inequality is stronger than (1.19).

The integrand on the right in (1.25) is

$$s_n^4(\omega) = \sum r_\alpha(\omega)r_\beta(\omega)r_\gamma(\omega)r_\delta(\omega), \tag{1.26}$$

where the four indices range independently from 1 to n. Depending on how the indices match up, each term in this sum reduces to one of the following five forms, where in each case the indices are now *distinct*:

$$\begin{cases} r_i^4(\omega) = 1, \\ r_i^2(\omega)r_j^2(\omega) = 1, \\ r_i^2(\omega)r_j(\omega)r_k(\omega) = r_j(\omega)r_k(\omega), \\ r_i^3(\omega)r_j(\omega) = r_i(\omega)r_j(\omega), \\ r_i(\omega)r_j(\omega)r_k(\omega)r_l(\omega). \end{cases} \tag{1.27}$$

If, for example, k exceeds i, j, and l, then the last product in (1.27) integrates to 0 over each dyadic interval of rank $k-1$, because $r_i(\omega)r_j(\omega)r_l(\omega)$ is constant there, while $r_k(\omega)$ is -1 on the left half and $+1$ on the right. Adding over the dyadic intervals of rank $k-1$ gives

$$\int_0^1 r_i(\omega)r_j(\omega)r_k(\omega)r_l(\omega)\,d\omega = 0.$$

This holds whenever the four indices are distinct. From this and (1.17) it follows that the last three forms in (1.27) integrate to 0 over the unit interval; of course, the first two forms integrate to 1.

The number of occurrences in the sum (1.26) of the first form in (1.27) is n. The number of occurrences of the second form is $3n(n-1)$, because there are n choices for the α in (1.26), three ways to match it with β, γ, or δ, and $n-1$ choices for the value common to the remaining two indices. A term-by-term integration of (1.26) therefore gives

$$\int_0^1 s_n^4(\omega)\,d\omega = n + 3n(n-1) \le 3n^2, \tag{1.28}$$

and it follows by (1.25) that

$$P\left[\omega: \left|\frac{1}{n}s_n(\omega)\right| \ge \epsilon\right] \le \frac{3}{n^2\epsilon^4}. \tag{1.29}$$

Fix a positive sequence $\{\epsilon_n\}$ going to 0 slowly enough that the series $\sum_n \epsilon_n^{-4}n^{-2}$ converges (take $\epsilon_n = n^{-1/8}$, for example). If $A_n = [\omega: |n^{-1}s_n(\omega)| \ge \epsilon_n]$, then $P(A_n) \le 3\epsilon_n^{-4}n^{-2}$ by (1.29), and so $\sum_n P(A_n) < \infty$.

If, for some m, ω lies in A_n^c for all n greater than or equal to m, then $|n^{-1}s_n(\omega)| < \epsilon_n$ for $n \ge m$, and it follows that ω is normal because $\epsilon_n \to 0$ (see (1.24)). In other words, for each m, $\bigcap_{n=m}^\infty A_n^c \subset N$, which is the same thing as $N^c \subset \bigcup_{n=m}^\infty A_n$. This last relation leads to the required covering: Given ϵ, choose m so that $\sum_{n=m}^\infty P(A_n) < \epsilon$. Now A_n is a finite disjoint union $\bigcup_k I_{nk}$ of intervals with $\sum_k |I_{nk}| = P(A_n)$, and therefore $\bigcup_{n=m}^\infty A_n$ is a countable union $\bigcup_{n=m}^\infty \bigcup_k I_{nk}$ of intervals (not disjoint, but that does not matter) with $\sum_{n=m}^\infty \sum_k |I_{nk}| = \sum_{n=m}^\infty P(A_n) < \epsilon$. The intervals I_{nk} $(n \ge m,\ k \ge 1)$ provide a covering of N^c of the kind the definition of negligibility calls for. ■

Strong Law Versus Weak

Theorem 1.2 is stronger than Theorem 1.1. A consideration of the forms of the two propositions will show that the strong law goes far beyond the weak law.

For each n let $f_n(\omega)$ be a step function on the unit interval, and consider the relation

$$\lim_{n\to\infty} P[\omega: |f_n(\omega)| \ge \epsilon] = 0 \tag{1.30}$$

together with the set

$$\left[\omega: \lim_{n\to\infty} f_n(\omega) = 0\right]. \tag{1.31}$$

If $f_n(\omega) = n^{-1}s_n(\omega)$, then (1.30) reduces to the weak law (1.15), and (1.31) coincides with the set (1.24) of normal numbers. According to a general result proved below (Theorem 5.2(ii)), whatever the step functions $f_n(\omega)$ may be, if the set (1.31) has negligible complement, then (1.30) holds for each positive ϵ. For this reason, a proof of Theorem 1.2 is automatically a proof of Theorem 1.1.

The converse, however, fails: There exist step functions $f_n(\omega)$ that satisfy (1.30) for each positive ϵ but for which (1.31) fails to have negligible complement (Example 5.4). For this reason, a proof of Theorem 1.1 is not automatically a proof of Theorem 1.2; the latter lies deeper and its proof is correspondingly more complex.

Length

According to Theorem 1.2, the complement N^c of the set of normal numbers is negligible. What if N itself were negligible? It would then follow that $(0,1] = N \cup N^c$ was negligible as well, which would disqualify negligibility as an explication of "practical impossibility," as a stand-in for "probability zero." The proof below of the "obvious" fact that an interval of positive length is not negligible (Theorem 1.3(ii)), while simple enough, does involve the most fundamental properties of the real number system.

Consider an interval $I=(a,b]$ of length $|I|=b-a$; see (1.1). Consider also a finite or infinite sequence of intervals $I_k=(a_k,b_k]$. While each of these intervals is bounded, they need not be subintervals of $(0,1]$.

Theorem 1.3. (i) *If* $\bigcup_k I_k \subset I$, *and the* I_k *are disjoint, then* $\Sigma_k |I_k| \le |I|$.

(ii) *If* $I \subset \bigcup_k I_k$ *(the* I_k *need not be disjoint), then* $|I| \le \Sigma_k |I_k|$.

(iii) *If* $I=\bigcup_k I_k$, *and the* I_k *are disjoint, then* $|I|=\Sigma_k |I_k|$.

PROOF. Of course (iii) follows from (i) and (ii).

PROOF OF (i): *Finite case.* Suppose there are n intervals. The result being obvious for $n=1$, assume that it holds for $n-1$. If a_n is the largest among $a_1,\ldots,a_n$ (this is just a matter of notation), then $\bigcup_{k=1}^{n-1}(a_k,b_k] \subset$ $(a,a_n]$, so that $\Sigma_{k=1}^{n-1}(b_k-a_k) \le a_n-a$ by the induction hypothesis, and hence $\Sigma_{k=1}^{n}(b_k-a_k) \le (a_n-a)+(b_n-a_n) \le b-a$.

Infinite case. If there are infinitely many intervals, each finite subcollection satisfies the hypotheses of (i), and so $\Sigma_{k=1}^{n}(b_k-a_k) \le b-a$ by the finite case. But as n is arbitrary, the result follows.

PROOF OF (ii): *Finite case.* Assume that the result holds for the case of $n-1$ intervals and that $(a,b] \subset \bigcup_{k=1}^{n}(a_k,b_k]$. Suppose that $a_n < b \le b_n$ (notation again). If $a_n \le a$, the result is obvious. Otherwise, $(a,a_n] \subset$ $\bigcup_{k=1}^{n-1}(a_k,b_k]$, so that $\Sigma_{k=1}^{n-1}(b_k-a_k) \ge a_n-a$ by the induction hypothesis and hence $\Sigma_{k=1}^{n}(b_k-a_k) \ge (a_n-a)+(b_n-a_n) \ge b-a$. The finite case thus follows by induction.

Infinite case. Suppose that $(a,b] \subset \bigcup_{k=1}^{\infty}(a_k,b_k]$. If $0<\epsilon<b-a$, the open intervals $(a_k,b_k+\epsilon 2^{-k})$ cover the closed interval $[a+\epsilon,b]$, and it follows by the Heine–Borel theorem [A13] that $[a+\epsilon,b] \subset \bigcup_{k=1}^{n}(a_k,b_k+\epsilon 2^{-k})$ for some n. But then $(a+\epsilon,b] \subset \bigcup_{k=1}^{n}(a_k,b_k+\epsilon 2^{-k}]$, and by the finite case, $b-(a+\epsilon) \le \Sigma_{k=1}^{n}(b_k+\epsilon 2^{-k}-a_k) \le \Sigma_{k=1}^{\infty}(b_k-a_k)+\epsilon$. Since ϵ was arbitrary, the result follows. ■

Theorem 1.3 will be the starting point for the theory of Lebesgue measure as developed in Sections 2 and 3. Taken together, parts (i) and (ii) of the theorem for only finitely many intervals I_k imply (1.4) for disjoint A and B. Like (1.4), they follow immediately from the additivity of the Riemann integral; but the point is to give an independent development of which the Riemann theory will be an eventual by-product.

To pass from the finite to the infinite case in part (i) of the theorem is easy. But to pass from the finite to the infinite case in part (ii) involves compactness, a profound idea underlying all of modern analysis. And it is part (ii) that shows that an interval I of positive length is not negligible: $|I|$ is

a positive lower bound for the sum of the lengths of the intervals in any covering of I.

The Measure Theory of Diophantine Approximation*

Diophantine approximation has to do with the approximation of real numbers x by rational fractions p/q. The measure theory of Diophantine approximation has to do with the degree of approximation that is possible if one disregards negligible sets of real x.

For each positive integer q, x must lie between some pair of successive multiples of $1/q$, so that for some p, $|x-p/q| \le 1/q$. Since for each q the intervals

$$\left(\frac{p}{q}-\frac{1}{2q}, \frac{p}{q}+\frac{1}{2q}\right] \tag{1.32}$$

decompose the line, the error of approximation can be further reduced to $1/2q$: For each q there is a p such that $|x-p/q| \le 1/2q$. These observations are of course trivial. But for "most" real numbers x there will be many values of p and q for which x lies very near the center of the interval (1.32), so that p/q is a very sharp approximation to x.

Theorem 1.4. *If x is irrational, there are infinitely many irreducible fractions p/q such that*

$$\left|x-\frac{p}{q}\right| < \frac{1}{q^2}. \tag{1.33}$$

This famous theorem of Dirichlet says that for infinitely many p and q, x lies in $(p/q-1/q^2, p/q+1/q^2)$ and hence is indeed very near the center of (1.32).

PROOF. For a positive integer Q, decompose $[0,1)$ into the Q subintervals $[(i-1)/Q, i/Q)$, $i=1,\ldots,Q$. The points (fractional parts) $\{qx\}=qx-\lfloor qx\rfloor$ for $q=0,1,\ldots,Q$ lie in $[0,1)$, and since there are $Q+1$ points† and only Q subintervals, it follows (Dirichlet's drawer principle) that some subinterval contains more than one point. Suppose that $\{q'x\}$ and $\{q''x\}$ lie in the same subinterval and $0 \le q' < q'' \le Q$. Take $q=q''-q'$ and $p=\lfloor q''x\rfloor-\lfloor q'x\rfloor$; then $1 \le q \le Q$ and $|qx-p|=|\{q''x\}-\{q'x\}| < 1/Q$:

$$\left|x-\frac{p}{q}\right| < \frac{1}{qQ} \le \frac{1}{q^2}. \tag{1.34}$$

If p and q have any common factors, cancel them; this will not change the left side of (1.34), and it will decrease q.

For each Q, therefore, there is an irreducible p/q satisfying (1.34).‡ Suppose there are only finitely many irreducible solutions of (1.33), say $p_1/q_1,\ldots,p_m/q_m$. Since x is irrational, the $|x-p_k/q_k|$ are all positive, and it is possible to choose Q so that Q^{-1} is smaller than each of them. But then the p/q of (1.34) is a solution of (1.33), and since $|x-p/q| < 1/Q$, there is a contradiction. ■

*This topic may be omitted.

†Although the fact is not technically necessary to the proof, these points are distinct: $\{q'x\}=\{q''x\}$ implies $(q''-q')x=\lfloor q''x\rfloor-\lfloor q'x\rfloor$, which in turn implies that x is rational unless $q'=q''$.

‡This much of the proof goes through even if x is rational.

In the measure theory of Diophantine approximation, one looks at the set of real x having such and such approximation properties and tries to show that this set is negligible or else that its complement is. Since the set of rationals is negligible, Theorem 1.4 implies such a result: Apart from a negligible set of x, (1.33) has infinitely many irreducible solutions.

What happens if the inequality (1.33) is tightened? Consider

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^2\varphi(q)}, \tag{1.35}$$

and let A_φ consist of the real x for which (1.35) has infinitely many irreducible solutions. Under what conditions on φ will A_φ have negligible complement? If $\varphi(q) \le 1$, then (1.35) is weaker than (1.33): $\varphi(q) > 1$ in the interesting cases. Since x satisfies (1.35) for infinitely many irreducible p/q if and only if $x - \lfloor x \rfloor$ does, A_φ may as well be redefined as the set of x in $(0,1)$ (or even as the set of irrational x in $(0,1)$) for which (1.35) has infinitely many solutions.

Theorem 1.5. *Suppose that φ is positive and nondecreasing. If*

$$\sum_q \frac{1}{q\varphi(q)} = \infty, \tag{1.36}$$

then A_φ has negligible complement.

Theorem 1.4 covers the case $\varphi(q) \equiv 1$. Although this is the natural place to state Theorem 1.5 in its general form, the proof, which involves continued fractions and the ergodic theorem, must be postponed; see Section 24, p. 324. The converse, on the other hand, has a very simple proof.

Theorem 1.6. *Suppose that φ is positive. If*

$$\sum_q \frac{1}{q\varphi(q)} < \infty, \tag{1.37}$$

then A_φ is negligible.

Proof. Given ϵ, choose q_0 so that $\sum_{q \ge q_0} 1/q\varphi(q) < \epsilon/4$. If $x \in A_\varphi$, then (1.35) holds for some $q \ge q_0$, and since $0 < x < 1$, the corresponding p lies in the range $0 \le p \le q$. Therefore,

$$A_\varphi \subset \bigcup_{q \ge q_0} \bigcup_{p=0}^{q} \left(\frac{p}{q} - \frac{1}{q^2\varphi(q)}, \frac{p}{q} + \frac{1}{q^2\varphi(q)} \right].$$

The right side here is a countable union of intervals covering A_φ, and the sum of their lengths is

$$\sum_{q \ge q_0} \sum_{p=0}^{q} \frac{2}{q^2\varphi(q)} = \sum_{q \ge q_0} \frac{2(q+1)}{q^2\varphi(q)} \le \sum_{q \ge q_0} \frac{4}{q\varphi(q)} < \epsilon.$$

Thus A_φ satisfies the definition ((1.22) and (1.23)) of negligibility. ■

If $\varphi_1(q) \equiv 1$, then (1.36) holds and hence A_{φ_1} has negligible complement (as follows also from Theorem 1.4). If $\varphi_2(q) = q^\epsilon$, however, then (1.37) holds and A_{φ_2} itself is negligible. Outside the negligible set $A_{\varphi_1}^c \cup A_{\varphi_2}$, therefore, $|x - p/q| < 1/q^2$ has infinitely many irreducible solutions but $|x - p/q| < 1/q^{2+\epsilon}$ has only finitely many. Similarly, since $\sum_q 1/(q \log q)$ diverges but $\sum_q 1/(q \log^{1+\epsilon} q)$ converges, outside a negligible set $|x - p/q| < 1/(q^2 \log q)$ has infinitely many irreducible solutions but $|x - p/q| < 1/(q^2 \log^{1+\epsilon} q)$ has only finitely many.

Rational approximations to x obtained by truncating its binary (or decimal) expansion are very inaccurate: see Example 4.17. The sharp rational approximations to x come from truncation of its continued-fraction expansion: see Section 24.

PROBLEMS

Some problems involve concepts not required for an understanding of the text, or concepts treated only in later sections; there are no problems whose solutions are used in the text itself. An arrow ↑ points back to a problem (the one immediately preceding if no number is given) the solution and terminology of which are assumed. See Notes on the Problems, p. 552.

1.1. (a) Show that a *discrete* probability space (see Example 2.8 for the formal definition) cannot contain an infinite sequence $A_1, A_2, \ldots$ of independent events each of probability $\frac{1}{2}$. Since A_n could be identified with heads on the nth toss of a coin, the existence of such a sequence would make this section superfluous.

(b) Suppose that $0 \le p_n \le 1$, and put $\alpha_n = \min\{p_n, 1 - p_n\}$. Show that, if $\sum_n \alpha_n$ diverges, then no discrete probability space can contain independent events $A_1, A_2, \ldots$ such that A_n has probability p_n.

1.2. Show that N and N^c are dense [A15] in $(0,1]$.

1.3. ↑ Define a set A to be *trifling*[†] if for each ϵ there exists a *finite* sequence of intervals I_k satisfying (1.22) and (1.23). This definition and the definition of negligibility apply as they stand to all sets on the real line, not just to subsets of $(0,1]$.

(a) Show that a trifling set is negligible.

(b) Show that the closure of a trifling set is also trifling.

(c) Find a bounded negligible set that is not trifling.

(d) Show that the closure of a negligible set may not be negligible.

(e) Show that finite unions of trifling sets are trifling but that this can fail for countable unions.

1.4. ↑ For $i = 0, \ldots, r-1$, let $A_r(i)$ be the set of numbers in $(0,1]$ whose nonterminating expansions in the base r do not contain the digit i.

(a) Show that $A_r(i)$ is trifling.

(b) Find a trifling set A such that every point in the unit interval can be represented in the form $x + y$ with x and y in A.

[†] Like *negligible*, *trifling* is a nonce word used only here. The trifling sets are exactly the sets of content 0: See Problem 3.15.

(c) Let $A_r(i_1,\ldots,i_k)$ consist of the numbers in the unit interval in whose base-r expansions the digits $i_1,\ldots,i_k$ nowhere appear consecutively in that order. Show that it is trifling. What does this imply about the monkey that types at random?

1.5. ↑ The *Cantor set* C can be defined as the closure of $A_3(1)$.

(a) Show that C is uncountable but trifling.

(b) From $[0,1]$ remove the open middle third $(\frac{1}{3},\frac{2}{3})$; from the remainder, a union of two closed intervals, remove the two open middle thirds $(\frac{1}{9},\frac{2}{9})$ and $(\frac{7}{9},\frac{8}{9})$. Show that C is what remains when this process is continued ad infinitum.

(c) Show that C is perfect [A15].

1.6. Put $M(t)=\int_0^1 e^{ts_n(\omega)}\,d\omega$, and show by successive differentiations under the integral that

$$M^{(k)}(0)=\int_0^1 s_n^k(\omega)\,d\omega. \tag{1.38}$$

Over each dyadic interval of rank n, $s_n(\omega)$ has a constant value of the form $\pm 1\pm 1\pm\cdots\pm 1$, and therefore $M(t)=2^{-n}\Sigma\exp t(\pm 1\pm 1\pm\cdots\pm 1)$, where the sum extends over all 2^n n-long sequences of $+1$'s and -1's. Thus

$$M(t)=\left(\frac{e^t+e^{-t}}{2}\right)^n=(\cosh t)^n. \tag{1.39}$$

Use this and (1.38) to give new proofs of (1.16), (1.18), and (1.28). (This, the method of moment generating functions, will be investigated systematically in Section 9.)

1.7. ↑ By an argument similar to that leading to (1.39) show that the Rademacher functions satisfy

$$\begin{aligned}\int_0^1 \exp\left[i\sum_{k=1}^n a_k r_k(\omega)\right]d\omega &= \prod_{k=1}^n \frac{e^{ia_k}+e^{-ia_k}}{2}\\ &= \prod_{k=1}^n \cos a_k.\end{aligned}$$

Take $a_k=t2^{-k}$, and from $\Sigma_{k=1}^\infty r_k(\omega)2^{-k}=2\omega-1$ deduce

$$\frac{\sin t}{t}=\prod_{k=1}^\infty \cos\frac{t}{2^k} \tag{1.40}$$

by letting $n\to\infty$ inside the integral above. Derive Vieta's formula

$$\frac{2}{\pi}=\frac{\sqrt{2}}{2}\,\frac{\sqrt{2+\sqrt{2}}}{2}\,\frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2}\cdots.$$

1.8. A number ω is normal in the base 2 if and only if for each positive ϵ there exists an $n_0(\epsilon,\omega)$ such that $|n^{-1}\Sigma_{i=1}^n d_i(\omega)-\frac{1}{2}|<\epsilon$ for all n exceeding $n_0(\epsilon,\omega)$.

Theorem 1.2 concerns the entire dyadic expansion, whereas Theorem 1.1 concerns only the beginning segment. Point up the difference by showing that for $\epsilon<\frac{1}{2}$ the $n_0(\epsilon,\omega)$ above cannot be the same for all ω in N—in other words, $n^{-1}\Sigma_{i=1}^n d_i(\omega)$ converges to $\frac{1}{2}$ for all ω in N, but not uniformly. But see Problem 13.9.

1.9. 1.3↑ (a) Using the finite form of Theorem 1.3(ii), together with Problem 1.3(b), show that a trifling set is nowhere dense [A15].

(b) Put $B=\bigcup_n(r_n-2^{-n-2},\,r_n+2^{-n-2}]$, where $r_1,r_2,\ldots$ is an enumeration of the rationals in $(0,1]$. Show that $(0,1]-B$ is nowhere dense but not trifling or even negligible.

(c) Show that a compact negligible set is trifling.

1.10. ↑ A set of the first category [A15] can be represented as a countable union of nowhere dense sets; this is a topological notion of smallness, just as negligibility is a metric notion of smallness. Neither condition implies the other:

(a) Show that the nonnegligible set N of normal numbers is of the first category by proving that $A_m=\bigcap_{n=m}^\infty[\omega\colon |n^{-1}s_n(\omega)|<\frac{1}{2}]$ is nowhere dense and $N\subset\bigcup_m A_m$.

(b) According to a famous theorem of Baire, a nonempty interval is *not* of the first category. Use this fact to prove that the negligible set $N^c=(0,1]-N$ is not of the first category.

1.11. Prove:

(a) If x is rational, (1.33) has only finitely many irreducible solutions.

(b) Suppose that $\varphi(q)\ge 1$ and (1.35) holds for infinitely many pairs p,q but only for finitely many relatively prime ones. Then x is rational.

(c) If φ goes to infinity too rapidly, then A_φ is negligible (Theorem 1.6). But however rapidly φ goes to infinity, A_φ is nonempty, even uncountable. *Hint:* Consider $x=\Sigma_{k=1}^\infty 1/2^{\alpha(k)}$ for integral $\alpha(k)$ increasing very rapidly to infinity.

SECTION 2. PROBABILITY MEASURES

Spaces

Let Ω be an arbitrary space or set of points ω. In probability theory Ω consists of all the possible results or outcomes ω of an experiment or observation. For observing the number of heads in n tosses of a coin the space Ω is $\{0,1,\ldots,n\}$; for describing the complete history of the n tosses Ω is the space of all 2^n n-long sequences of H's and T's; for an infinite sequence of tosses Ω can be taken as the unit interval as in the preceding section; for the number of α-particles emitted by a substance during a unit interval of time or for the number of telephone calls arriving at an exchange Ω is $\{0,1,2,\ldots\}$; for the position of a particle Ω is three-dimensional Euclidean space; for describing the motion of the particle Ω is an appropriate space of functions; and so on. Most Ω's to be considered are interesting from the point of view of geometry and analysis as well as that of probability.