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.

[†]For the discrete theory, presupposed here, see for example the first half of Volume 1 of FELLER. (Names in capital letters refer to the bibliography on p. 581.)

Probability and Measure, Anniversary Edition. Patrick Billingsley.

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 $|I|$:

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

If

$$A = \bigcup_{i=1}^{n} I_i = \bigcup_{i=1}^{n} (a_1, b_1], \tag{1.2}$$

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

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

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

[†]A notation [An] refers to paragraph n of the appendix beginning on p. 571; this is a collection of mathematical definitions and facts required in the text.

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.

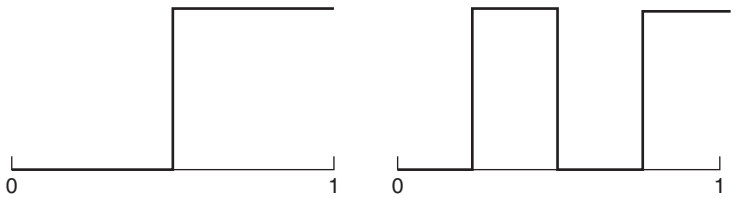
The definition (1.3) is also the starting point for a mathematical representation of an infinite sequence of tosses of a coin. With each ω associate its nonterminating dyadic expansion

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

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 $.111\ldots$.



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 l's. Such an ω satisfies

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

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

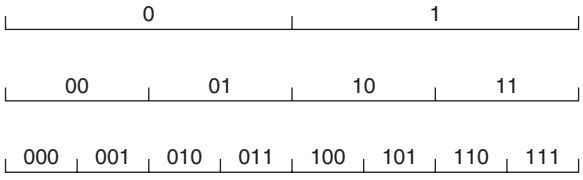
where the extreme values of ω correspond to the case $d_i(\omega) = 0$ for $i > n$ and the case $d_i(\omega) = 1$ for $i > n$. The second case can be achieved, but since the binary expansions represented by the $d_i(\omega)$ are nonterminating—do not end in 0's—the first cannot, and ω must actually exceed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

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

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),

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

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 $[\omega\colon d_{n+1}(\omega) = u]$ is thus a disjoint union of 2^n intervals of length $1/2^{n+1}$ and hence has probability $\frac{1}{2}$: $P[\omega\colon d_n(\omega) = u] = \frac{1}{2}$ 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

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

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, ...), 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, ...)—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,[†]

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

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.

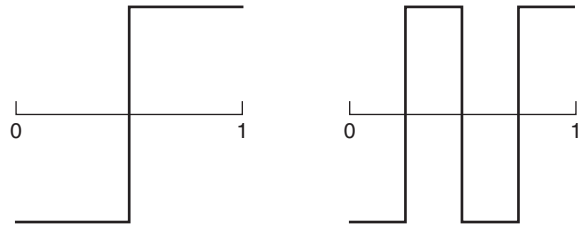
Since $d_i(\omega)$ is constant over each dyadic interval of rank n if $i \leq n$, the sum $\sum_{i=1}^{n} d_i(\omega)$ is also constant over each dyadic interval of rank n. The set in (1.12) is therefore the union of certain of the intervals (1.9), and so its probability is well defined by (1.3).

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

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

simpler if the $d_n(\omega)$ are replaced by the *Rademacher functions*,

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



Graph of $r_1(\omega)$ Graph of $r_2(\omega)$

Consider the partial sums

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

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

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

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

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

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 $r_i(\omega)$ 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

$$\int_0^1 r_i(\omega)r_j(\omega)\,d\omega = 0, \quad i \neq j. \tag{1.17}$$

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

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

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

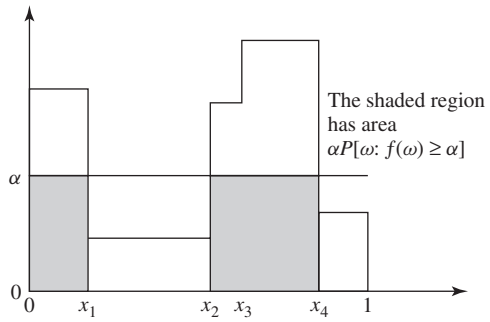
$$P[\omega\colon |s_n(\omega)| \geq n\epsilon] \leq \frac{1}{n^2\epsilon^2}\int_0^1 s_n^2(\omega)\,d\omega = \frac{1}{n\epsilon^2}. \tag{1.19}$$

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 $[\omega\colon f(\omega) \geq \alpha]$ is for $\alpha > 0$ a finite union of intervals and*

$$P[\omega\colon f(\omega) \geq \alpha] \leq \frac{1}{\alpha}\int_0^1 f(\omega)\,d\omega. \tag{1.20}$$



Proof. The set in question is the union of the intervals $(x_{j-1}, x_j]$ for which $c_j \geq \alpha$. If $\sum'$ denotes summation over those j satisfying $c_j \geq \alpha$, then

$P[\omega\colon f(\omega) \geq \alpha] = \sum'(x_j - x_{j-1})$ by the definition (1.3). On the other hand, since the c_j are all nonnegative by hypothesis, (1.6) gives

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

Hence (1.20). ■

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

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

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

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

[†]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*.

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, ...). 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 $(u_3, u_6, \ldots)$

[†]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.

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

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

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.

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

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

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

and it follows by (1.25) that

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

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 $A_n = [\omega\colon |n^{-1}s_n(\omega)| \geq \epsilon_n]$, then $P(A_n) \leq 3\epsilon_n^{-4}n^{-2}$ 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 $n \geq m$, 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 $I_{nk} (n \geq m, k \geq 1)$ 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

$$\lim_{n \leftarrow \infty} P[\omega\colon |f_n(\omega)| \geq \epsilon] = 0 \tag{1.30}$$

together with the set

$$[\omega\colon \lim_{n \leftarrow \infty} f_n(\omega) = 0]. \tag{1.31}$$

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 (1.30) 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 $\sum_k |I_k| \le |I|$.

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

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

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 $\sum_{k=1}^{n-1}(b_k - a_k) \le a_n - a$ by the induction hypothesis, and hence $\sum_{k=1}^{n}(b_k - a_k) \le (a_n - a) + (b_n - a_n) \le b - a$.

Infinite case. If there are infinitely many intervals, each finite subcollection satisfies the hypotheses of (i), and so $\sum_{k=1}^{n}(b_k - a_k) \le b - a$ 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 $\sum_{k=1}^{n-1}(b_k - a_k) \ge a_n - a$ by the induction hypothesis and hence $\sum_{k=1}^{n}(b_k - a_k) \ge (a_n - a) + (b_n - a_n) \ge b - a$. 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, $b - (a + \epsilon) \le \sum_{k=1}^{n}(b_k + \epsilon 2^{-k} - a_k) \le \sum_{k=1}^{\infty}(b_k - a_k) + \epsilon$. 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) $\{qx\} = qx - \lfloor qx \rfloor$

†This topic may be omitted.

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. ■

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.

[†]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.

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 $\sum_{q \geq q_0} 1/q\varphi(q) < \epsilon/4$. If $x \in A_\varphi$, then (1.35) holds for some $q \geq q_0$, and since $0 < x < 1$, the corresponding p lies in the range $0 \leq p \leq q$. Therefore,

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

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

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

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 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. 589.

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.

(c) Let $A_r(i_l, \ldots, i_k)$ consist of the numbers in the unit interval in whose base-r expansions the digits $i_l, \ldots, i_k$ 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 $\left(\frac{1}{3}, \frac{2}{3}\right)$; from the remainder, a union of two closed intervals, remove the two open middle thirds $\left(\frac{1}{9}, \frac{2}{9}\right)$ and $\left(\frac{7}{9}, \frac{8}{9}\right)$. Show that C is what remains when this process is continued ad infinitum.

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

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

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 $M(t) = 2^{-n} \sum \exp t(\pm 1 \pm 1 \pm \cdots \pm 1)$, 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

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

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

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

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 $|n^{-1} \sum_{i=1}^n d_i(\omega) - \frac{1}{2}| < \epsilon$ 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, $n^{-1} \sum_{i=1}^n d_i(\omega)$ 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 $A_m = \bigcap_{n=m}^{\infty} \left[\omega\colon |n^{-1}s_n(\omega)| < \frac{1}{2}\right]$ 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 $\varphi(q) \geq 1$ 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 $x = \sum_{k=1}^{\infty} 1/2^{\alpha(k)}$ 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.

Viewed probabilistically, a subset of Ω is an *event* and an element ω of Ω is a *sample point*.

Assigning Probabilities

In setting up a space Ω as a probabilistic model, it is natural to try and assign probabilities to as many events as possible. Consider again the case $\Omega = (0, 1]$—the unit interval. It is natural to try and go beyond the definition (1.3) and assign probabilities in a systematic way to sets other than finite unions of intervals. Since the set of nonnormal numbers is negligible, for example, one feels it ought to have probability 0. For another probabilistically interesting set that is not a finite union of intervals, consider

$$\bigcup_{n=1}^{\infty}[\omega:\ -a < s_1(\omega), \ldots, s_{n-1}(\omega) < b, s_n(\omega) = -a], \tag{2.1}$$

where a and b are positive integers. This is the event that the gambler's fortune reaches $-a$ before it reaches $+b$; it represents ruin for a gambler with a dollars playing against an adversary with b dollars, the rule being that they play until one or the other runs out of capital.

The union in (2.1) is countable and disjoint, and for each n the set in the union is itself a union of certain of the intervals (1.9). Thus (2.1) is a countably infinite disjoint union of intervals, and it is natural to take as its probability the sum of the lengths of these constituent intervals. Since the set of normal numbers is not a countable disjoint union of intervals, however, this extension of the definition of probability would still not cover all the interesting sets (events) in (0, 1].

It is, in fact, not fruitful to try to predict just which sets probabilistic analysis will require and then assign probabilities to them in some *ad hoc* way. The successful procedure is to develop a general theory that assigns probabilities at once to the sets of a class so extensive that most of its members never actually arise in probability theory. That being so, why not ask for a theory that goes all the way and applies to *every* set in a space Ω? In the case of the unit interval, should there not exist a well-defined probability that the random point ω lies in A, whatever the set A may be? The answer turns out to be no (see p. 45), and it is necessary to work within subclasses of the class of all subsets of a space Ω. The classes of the appropriate kinds—the fields and σ-fields—are defined and studied in this section. The theory developed here covers the spaces listed above, including the unit interval, and a great variety of others.

Classes of Sets

It is necessary to single out for special treatment classes of subsets of a space Ω, and to be useful, such a class must be closed under various of the operations of set theory. Once again the unit interval provides an instructive example.

EXAMPLE 2.1[†]

Consider the set N of normal numbers in the form (1.24), where $s_n(\omega)$ is the sum of the first n Rademacher functions. Since a point ω lies in N if and only if $\lim_n n^{-1}s_n(\omega) = 0, N$ can be put in the form

$$N = \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} [\omega\colon |n^{-1}s_n(\omega)| < k^{-1}]. \tag{2.2}$$

Indeed, because of the very meaning of union and of intersection, ω lies in the set on the right here if and only if for every k there exists an m such that $|n^{-1}s_n(\omega)| < k^{-1}$ holds for all $n \geq m$, and this is just the definition of convergence to 0—with the usual ϵ replaced by k^{-1} to avoid the formation of an uncountable intersection. Since $s_n(\omega)$ is constant over each dyadic interval of rank n, the set $[\omega\colon n^{-1}s_n(\omega)| < k^{-1}]$ is a finite disjoint union of intervals. The formula (2.2) shows explicitly how N is constructed in steps from these simpler sets.

A systematic treatment of the ideas in Section 1 thus requires a class of sets that contains the intervals and is closed under the formation of countable unions and intersections. Note that a singleton [A1] $\{x\}$ is a countable intersection $\cap_n (x - n^{-1}, x]$ of intervals. If a class contains all the singletons and is closed under the formation of *arbitrary* unions, then of course it contains *all* the subsets of Ω. As the theory of this section and the next does not apply to such extensive classes of sets, attention must be restricted to countable set-theoretic operations and in some cases even to finite ones.

Consider now a completely arbitrary nonempty space Ω. A class $\mathscr{F}$ of subsets of Ω is called a *field*[‡] if it contains Ω itself and is closed under the formation of complements and finite unions:

(i) $\Omega \in \mathscr{F}$;

(ii) $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$;

(iii) $A, B \in \mathscr{F}$ implies $A \cup B \in \mathscr{F}$.

Since Ω and the empty set $\emptyset$ are complementary, (i) is the same in the presence of (ii) as the assumption $\emptyset \in \mathscr{F}$. In fact, (i) simply ensures that $\mathscr{F}$ is nonempty: If $A \in \mathscr{F}$, then $A^c \in \mathscr{F}$ by (ii) and $\Omega = A \cup A^c \in \mathscr{F}$ by (iii).

By DeMorgan's law, $A \cap B = (A^c \cup B^c)^c$ and $A \cup B = (A^c \cap B^c)^c$. If $\mathscr{F}$ is closed under complementation, therefore, it is closed under the formation of

[†]Many of the examples in the book simply illustrate the concepts at hand, but others contain definitions and facts needed subsequently.

[‡]The term *algebra* is often used in place of *field*.

finite unions if and only if it is closed under the formation of finite intersections. Thus (iii) can be replaced by the requirement

(iii′) $A, B \in \mathscr{F}$ implies $A \cap B \in \mathscr{F}$.

A class $\mathscr{F}$ of subsets of Ω is a *σ-field* if it is a field and if it is also closed under the formation of *countable* unions:

(iv) $A_1, A_2, \ldots \in \mathscr{F}$ implies $A_1 \cup A_2 \cup \cdots \in \mathscr{F}$.

By the infinite form of DeMorgan's law, assuming (iv) is the same thing as assuming

(iv′) $A_1, A_2, \ldots \in \mathscr{F}$ implies $A_1 \cap A_2 \cap \cdots \in \mathscr{F}$.

Note that (iv) implies (iii) because one can take $A_1 = A$ and $A_n = B$ for $n \geq 2$. A field is sometimes called a *finitely additive* field to stress that it need not be a σ-field. A set in a given class $\mathscr{F}$ is said to be *measurable* $\mathscr{F}$ or to be an *$\mathscr{F}$-set*. A field or σ-field of subsets of Ω will sometimes be called a field or σ-field *in* Ω.

EXAMPLE 2.2

Section 1 began with a consideration of the sets (1.2), the finite disjoint unions of subintervals of $\Omega = (0, 1]$. Augmented by the empty set, this class is a field $\mathscr{B}_0$: Suppose that $A = (a_1, a_1'] \cup \cdots \cup (a_m, a_m']$, where the notation is so chosen that $a_1 \leq \cdots \leq a_m$. If the $(a_i, a_i']$ are disjoint, then A^c is $(0, a_1] \cup (a_1', a_2] \cup \cdots \cup (a_{m-1}', a_m] \cup (a_m', 1]$ and so lies in $\mathscr{B}_0$ (some of these intervals may be empty, as a_i' and a_{i+1} may coincide). If $B = (b_1, b_1'] \cup \cdots \cup (b_n, b_n']$, the $(b_j, b_j']$ again disjoint, then $A \cap B = \bigcup_{i=1}^m \bigcup_{j=1}^n \{(a_i, a_i'] \cap (b_j, b_j']\}$; each intersection here is again an interval or else the empty set, and the union is disjoint, and hence $A \cap B$ is in $\mathscr{B}_0$. Thus $\mathscr{B}_0$ satisfies (i), (ii), and (iii′).

Although $\mathscr{B}_0$ is a field, it is not a σ-field: It does not contain the singletons $\{x\}$, even though each is a countable intersection $\bigcap_n (x - n^{-1}, x]$ of $\mathscr{B}_0$-sets. And $\mathscr{B}_0$ does not contain the set (2.1), a countable union of intervals that cannot be represented as a finite union of intervals. The set (2.2) of normal numbers is also outside $\mathscr{B}_0$.

The definitions above involve distinctions perhaps most easily made clear by a pair of artificial examples.

EXAMPLE 2.3

Let $\mathscr{F}$ consist of the finite and the cofinite sets (A being cofinite if A^c is finite). Then $\mathscr{F}$ is a field. If Ω is finite, then $\mathscr{F}$ contains all the subsets of Ω and hence is a σ-field as well. If Ω is infinite, however, then $\mathscr{F}$ is not a

σ-field. Indeed, choose in Ω a set A that is countably infinite and has infinite complement. (For example, choose a sequence $\omega_1, \omega_2, \ldots$ of distinct points in Ω and take $A = \{\omega_2, \omega_4, \ldots\}$.) Then $A \notin \mathscr{F}$, even though A is the union, necessarily countable, of the singletons it contains and each singleton is in $\mathscr{F}$. This shows that the definition of σ-field is indeed more restrictive than that of field.

EXAMPLE 2.4

Let $\mathscr{F}$ consist of the countable and the cocountable sets (A being cocountable if A^c is countable). Then $\mathscr{F}$ is a σ-field. If Ω is uncountable, then it contains a set A such that A and A^c are both uncountable.[†] Such a set is not in $\mathscr{F}$, which shows that even a σ-field may not contain all the subsets of Ω; furthermore, this set is the union (uncountable) of the singletons it contains and each singleton is in $\mathscr{F}$, which shows that a σ-field may not be closed under the formation of arbitrary unions.

The largest σ-field in Ω is the *power class* 2^Ω, consisting of *all* the subsets of Ω; the smallest σ-field consists only of the empty set and Ω itself.

The elementary facts about fields and σ-fields are easy to prove: If $\mathscr{F}$ is a field, then $A, B \in \mathscr{F}$ implies $A - B = A \cap B^c \in \mathscr{F}$ and $A \triangle B = (A - B) \cup (B - A) \in \mathscr{F}$. Further, it follows by induction on n that $A_1, \ldots, A_n \in \mathscr{F}$ implies $A_1 \cup \cdots \cup A_n \in \mathscr{F}$ and $A_1 \cap \cdots \cap A_n \in \mathscr{F}$.

A field is closed under the finite set-theoretic operations, and a σ-field is closed also under the countable ones. The analysis of a probability problem usually begins with the sets of some rather small class $\mathscr{A}$, such as the class of subintervals of (0, 1]. As in Example 2.1, probabilistically natural constructions involving finite and countable operations can then lead to sets outside the initial class $\mathscr{A}$. This leads one to consider a class of sets that (i) contains $\mathscr{A}$ and (ii) is a σ-field; it is natural and convenient, as it turns out, to consider a class that has these two properties and that in addition (iii) is in a certain sense as small as possible. As will be shown, this class is the *intersection of all the σ-fields containing $\mathscr{A}$*; it is called the *σ-field generated by $\mathscr{A}$* and is denoted by $\sigma(\mathscr{A})$.

There do exist σ-fields containing $\mathscr{A}$, the class of all subsets of Ω being one. Moreover, a completely arbitrary intersection of σ-fields (however many of them there may be) is itself a σ-field: Suppose that $\mathscr{F} = \bigcap_\theta \mathscr{F}_\theta$, where θ

[†]If Ω is the unit interval, for example, take $A = (0, \frac{1}{2}]$, say. To show that the general uncountable Ω contains such an A requires the axiom of choice [A8]. As a matter of fact, to prove the existence of the sequence alluded to in Example 2.3 requires a form of the axiom of choice, as does even something so apparently down-to-earth as proving that a countable union of negligible sets is negligible. Most of us use the axiom of choice completely unaware of the fact. Even Borel and Lebesgue did; see WAGON, pp. 217 ff.

ranges over an arbitrary index set and each $\mathscr{F}_\theta$ is a σ-field. Then $\Omega \in \mathscr{F}_\theta$ for all θ, so that $\Omega \in \mathscr{F}$. And $A \in \mathscr{F}$ implies for each θ that $A \in \mathscr{F}_\theta$ and hence $A^c \in \mathscr{F}_\theta$, so that $A^c \in \mathscr{F}$. If $A_n \in \mathscr{F}$ for each n, then $A_n \in \mathscr{F}_\theta$ for each n and θ, so that $\bigcup_n A_n$ lies in each $\mathscr{F}_\theta$ and hence in $\mathscr{F}$.

Thus the intersection in the definition of $\sigma(\mathscr{A})$ is indeed a σ-field containing $\mathscr{A}$. It is as small as possible, in the sense that it is contained in every σ-field that contains $\mathscr{A}$: if $\mathscr{A} \subset \mathscr{G}$ and $\mathscr{G}$ is a σ-field, then $\mathscr{G}$ is one of the σ-fields in the intersection defining $\sigma(\mathscr{A})$, so that $\sigma(\mathscr{A}) \subset \mathscr{G}$. Thus $\sigma(\mathscr{A})$ has these three properties:

(i) $\mathscr{A} \subset \sigma(\mathscr{A})$;

(ii) $\sigma(\mathscr{A})$ *is a* σ*-field*;

(iii) *if* $\mathscr{A} \subset \mathscr{G}$ *and* $\mathscr{G}$ *is a* σ*-field, then* $\sigma(\mathscr{A}) \subset \mathscr{G}$.

The importance of σ-fields will gradually become clear.

EXAMPLE 2.5

If $\mathscr{F}$ is a σ-field, then obviously $\sigma(\mathscr{F}) = \mathscr{F}$. If $\mathscr{A}$ consists of the singletons, then $\sigma(\mathscr{A})$ is the σ-field in Example 2.4. If $\mathscr{A}$ is empty or $\mathscr{A} = \{\emptyset\}$ or $\mathscr{A} = \{\Omega\}$, then $\sigma(\mathscr{A}) = \{\emptyset, \Omega\}$. If $\mathscr{A} \subset \mathscr{A}'$, then $\sigma(\mathscr{A}) \subset \sigma(\mathscr{A}')$. If $\mathscr{A} \subset \mathscr{A}' \subset \sigma(\mathscr{A})$, then $\sigma(\mathscr{A}) = \sigma(\mathscr{A}')$.

EXAMPLE 2.6

Let $\mathscr{I}$ be the class of subintervals of $\Omega = (0, 1]$, and define $\mathscr{B} = \sigma(\mathscr{I})$. The elements of $\mathscr{B}$ are called the *Borel sets* of the unit interval. The field $\mathscr{B}_0$ of Example 2.2 satisfies $\mathscr{I} \subset \mathscr{B}_0 \subset \mathscr{B}$, and hence $\sigma(\mathscr{B}_0) = \mathscr{B}$.

Since $\mathscr{B}$ contains the intervals and is a σ-field, repeated finite and countable set-theoretic operations starting from intervals will never lead outside $\mathscr{B}$. Thus $\mathscr{B}$ contains the set (2.2) of normal numbers. It also contains for example the open sets in (0, 1]: If G is open and $x \in G$, then there exist rationals a_x and b_x such that $x \in (a_x, b_x] \subset G$. But then $G = \bigcup_{x \in G} (a_x, b_x]$; since there are only countably many intervals with rational endpoints, G is a *countable* union of elements of $\mathscr{I}$ and hence lies in $\mathscr{B}$.

In fact, $\mathscr{B}$ contains all the subsets of (0, 1] actually encountered in ordinary analysis and probability. It is large enough for all "practical" purposes. It does not contain every subset of the unit interval, however; see the end of Section 3 (p. 45). The class $\mathscr{B}$ will play a fundamental role in all that follows.

Probability Measures

A *set function* is a real-valued function defined on some class of subsets of Ω. A set function P on a field $\mathscr{F}$ is a *probability measure* if it satisfies these conditions:

(i) $0 \le P(A) \le 1$ for $A \in \mathscr{F}$;

(ii) $P(\emptyset) = 0, P(\Omega) = 1$;

(iii) if $A_1, A_2, \ldots$ is a disjoint sequence of $\mathscr{F}$-sets and if $\bigcup_{k=1}^{\infty} A_k \in \mathscr{F}$, then[†]

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k). \tag{2.3}$$

The condition imposed on the set function P by (iii) is called *countable additivity*. Note that, since $\mathscr{F}$ is a field but perhaps not a σ-field, it is necessary in (iii) to assume that $\bigcup_{k=1}^{\infty} A_k$ lies in $\mathscr{F}$. If $A_1, \ldots, A_n$ are disjoint $\mathscr{F}$-sets, then $\bigcup_{k=1}^{n} A_k$ is also in $\mathscr{F}$ and (2.3) with $A_{n+1} = A_{n+2} = \cdots = \emptyset$ gives

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k). \tag{2.4}$$

The condition that (2.4) holds for disjoint $\mathscr{F}$-sets is *finite additivity*; it is a consequence of countable additivity. It follows by induction on n that P is finitely additive if (2.4) holds for $n = 2$—if $P(A \cup B) = P(A) + P(B)$ for disjoint $\mathscr{F}$-sets A and B.

The conditions above are redundant, because (i) can be replaced by $P(A) \ge 0$ and (ii) by $P(\Omega) = 1$. Indeed, the weakened forms (together with (iii)) imply that $P(\Omega) = P(\Omega) + P(\emptyset) + P(\emptyset) + \cdots$, so that $P(\emptyset) = 0$, and $1 = P(\Omega) = P(A) + P(A^c)$, so that $P(A) \le 1$.

EXAMPLE 2.7

Consider as in Example 2.2 the field $\mathscr{B}_0$ of finite disjoint unions of subintervals of $\Omega = (0, 1]$. The definition (1.3) assigns to each $\mathscr{B}_0$-set a number—the sum of the lengths of the constituent intervals—and hence specifies a set function P on $\mathscr{B}_0$. Extended inductively, (1.4) says that P is finitely additive. In Section 1 this property was deduced from the additivity of the Riemann integral (see (1.5)). In Theorem 2.2 below, the finite additivity of P will be proved from first principles, and it will be shown that P is, in fact, countably additive—is a probability measure on the field $\mathscr{B}_0$. The hard part of the argument is in the proof of Theorem 1.3, already done; the rest will be easy.

[†]As the left side of (2.3) is invariant under permutations of the A_n, the same must be true of the right side. But in fact, according to Dirichlet's theorem [A26], a nonnegative series has the same value whatever order the terms are summed in.

If $\mathscr{F}$ is a σ-field in Ω and P is a probability measure on $\mathscr{F}$, the triple $(\Omega, \mathscr{F}, P)$ is called a *probability measure space*, or simply a *probability space*. A *support* of P is any $\mathscr{F}$-set A for which $P(A) = 1$.

EXAMPLE 2.8

Let $\mathscr{F}$ be the σ-field of all subsets of a countable space Ω, and let $p(\omega)$ be a nonnegative function on Ω. Suppose that $\sum_{\omega\in\Omega} p(\omega) = 1$, and define $P(A) = \sum_{\omega\in A} p(\omega)$; since $p(\omega) \geq 0$, the order of summation is irrelevant by Dirichlet's theorem [A26]. Suppose that $A = \bigcup_{i=1}^{\infty} A_i$, where the A_i are disjoint, and let $\omega_{i1}, \omega_{i2}, \ldots$ be the points in A_i. By the theorem on nonnegative double series [A27], $P(A) = \sum_{ij} p(\omega_{ij}) = \sum_i \sum_j p(\omega_{ij}) = \sum_i P(A_i)$, and so P is countably additive. This $(\Omega, \mathscr{F}, P)$ is a *discrete probability space*. It is the formal basis for discrete probability theory.

EXAMPLE 2.9

Now consider a probability measure P on an arbitrary σ-field $\mathscr{F}$ in an arbitrary space Ω; P is a *discrete probability measure* if there exist finitely or countably many points ω_k and masses m_k such that $P(A) = \sum_{\omega_k\in A} m_k$ for A in $\mathscr{F}$. Here P is discrete, but the space itself may not be. In terms of indicator functions, the defining condition is $P(A) = \sum_k m_k I_A(\omega_k)$ for $A \in \mathscr{F}$. If the set $\{\omega_1, \omega_2, \ldots\}$ lies in $\mathscr{F}$, then it is a support of P.

If there is just one of these points, say ω_0, with mass $m_0 = 1$, then P is a *unit mass* at ω_0. In this case $P(A) = I_A(\omega_0)$ for $A \in \mathscr{F}$.

Suppose that P is a probability measure on a field $\mathscr{F}$, and that $A, B \in \mathscr{F}$ and $A \subset B$. since $P(A) + P(B - A) = P(B)$, *P is monotone:*

$$P(A) \leq P(B) \quad \text{if } A \subset B. \tag{2.5}$$

It follows further that $P(B - A) = P(B) - P(A)$, and as a special case,

$$P(A^c) = 1 - P(A). \tag{2.6}$$

Other formulas familiar from the discrete theory are easily proved. For example,

$$P(A) + P(B) = P(A \cup B) + P(A \cap B), \tag{2.7}$$

the common value of the two sides being $P(A \cup B^c) + 2P(A \cap B) + P(A^c \cap B)$. Subtraction gives

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.8}$$

This is the case $n = 2$ of the general *inclusion-exclusion formula*:

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_l \cap \cdots \cap A_n). \tag{2.9}$$

To deduce this inductively from (2.8), note that (2.8) gives

$$P\left(\bigcup_{k=1}^{n+1} A_k\right) = P\left(\bigcup_{k=1}^{n} A_k\right) + P(A_{n+1}) - P\left(\bigcup_{k=1}^{n} (A_k \cap A_{n+1})\right).$$

Applying (2.9) to the first and third terms on the right gives (2.9) with $n+1$ in place of n.

If $B_1 = A_1$ and $B_k = A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c$, then the B_k are disjoint and $\bigcup_{k=1}^n A_k = \bigcup_{k=1}^n B_k$, so that $P(\cup_{k=1}^n A_k) = \sum_{k=1}^n P(B_k)$. Since $P(B_k) \le P(A_k)$ by monotonicity, this establishes the *finite subadditivity* of P:

$$P\left(\bigcup_{k=1}^{n} A_k\right) \le \sum_{k=1}^{n} P(A_k). \tag{2.10}$$

Here, of course, the A_k need not be disjoint. Sometimes (2.10) is called *Boole's inequality*.

In these formulas all the sets are naturally assumed to lie in the field $\mathcal{F}$. The derivations above involve only the finite additivity of P. Countable additivity gives further properties:

THEOREM 2.1

Let P be a probability measure on a field $\mathcal{F}$.

(i) *Continuity from below: If A_n and A lie in $\mathcal{F}$ and*[†] *$A_n \uparrow A$, then $P(A_n) \uparrow P(A)$.*

(ii) *Continuity from above: If A_n and A lie in $\mathcal{F}$ and $A_n \downarrow A$, then $P(A_n) \downarrow P(A)$.*

(iii) *Countable subadditivity: If $A_1, A_2, \ldots$ and $\bigcup_{k=1}^{\infty} A_k$ lie in $\mathcal{F}$ (the A_k need not be disjoint), then*

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) \le \sum_{k=1}^{\infty} P(A_k). \tag{2.11}$$

[†]For the notation, see [A4] and [A10].

Proof. For (i), put $B_1 = A_1$ and $B_k = A_k - A_{k-1}$. Then the B_k are disjoint, $A = \bigcup_{k=1}^{\infty} B_k$, and $A_n = \bigcup_{k=1}^{n} B_k$, so that by countable and finite additivity, $P(A) = \sum_{k=1}^{\infty} P(B_k) = \lim_n \sum_{k=1}^{n} P(B_k) = \lim_n P(A_n)$. For (ii), observe that $A_n \downarrow A$ implies $A_n^c \uparrow A^c$, so that $1 - P(A_n) \uparrow 1 - P(A)$.

As for (iii), increase the right side of (2.10) to $\sum_{k=1}^{\infty} P(A_k)$ and then apply part (i) to the left side. ■

EXAMPLE 2.10

In the presence of finite additivity, a special case of (ii) implies countable additivity. *If P is a finitely additive probability measure on the field $\mathcal{F}$, and if $A_n \downarrow \emptyset$ for sets A_n in $\mathcal{F}$ implies $P(A_n) \downarrow 0$, then P is countably additive.* Indeed, if $B = \bigcup_k B_k$ for disjoint sets B_k (B and the B_k in $\mathcal{F}$), then $C_n = \bigcup_{k>n} B_k = B - \bigcup_{k \le n} B_k$ lies in the field $\mathcal{F}$, and $C_n \downarrow \emptyset$. The hypothesis, together with finite additivity, gives $P(B) - \sum_{k=1}^{n} P(B_k) = P(C_n) \to 0$, and hence $P(B) = \sum_{k=1}^{\infty} P(B_k)$.

Lebesgue Measure on the Unit Interval

The definition (1.3) specifies a set function on the field $\mathcal{B}_0$ of finite disjoint unions of intervals in (0, 1]; the problem is to prove P countably additive. It will be convenient to change notation from P to λ, and to denote by $\mathcal{I}$ the class of subintervals $(a, b]$ of (0, 1]; then $\lambda(I) = |I| = b - a$ is ordinary length. Regard $\emptyset$ as an element of $\mathcal{I}$ of length 0. If $A = \bigcup_{i=1}^{n} I_i$, the I_i being disjoint $\mathcal{I}$-sets, the definition (1.3) in the new notation is

$$\lambda(A) = \sum_{i=1}^{n} \lambda(I_i) = \sum_{i=1}^{n} |I_i|. \tag{2.12}$$

As pointed out in Section 1, there is a question of uniqueness here, because A will have other representations as a finite disjoint union $\bigcup_{j=1}^{m} J_j$ of $\mathcal{I}$-sets. But $\mathcal{I}$ is closed under the formation of finite intersections, and so the finite form of Theorem 1.3(iii) gives

$$\sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} \sum_{j=1}^{m} |I_i \cap J_j| = \sum_{j=1}^{m} |J_j|. \tag{2.13}$$

(Some of the $I_i \cap J_j$ may be empty, but the corresponding lengths are then 0.) The definition is indeed consistent.

Thus (2.12) defines a set function λ on $\mathcal{B}_0$, a set function called *Lebesgue measure*.

THEOREM 2.2

Lebesgue measure λ *is a* (*countably additive*) *probability measure on the field* $\mathscr{B}_0$.

Proof. Suppose that $A = \bigcup_{k=1}^{\infty} A_k$, where A and the A_k are $\mathscr{B}_0$-sets and the A_k are disjoint. Then $A = \bigcup_{i=1}^{n} I_i$ and $A_k = \bigcup_{j=1}^{m_k} J_{kj}$ are disjoint unions of $\mathscr{I}$-sets, and (2.12) and Theorem 1.3(iii) give

$$\lambda(A) = \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |I_i \cap J_{kj}| \tag{2.14}$$

$$= \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |J_{kj}| = \sum_{k=1}^{\infty} \lambda(A_k).$$

■

In Section 3 it is shown how to extend λ from $\mathscr{B}_0$ to the larger class $\mathscr{B} = \sigma(\mathscr{B}_0)$ of Borel sets in (0, 1]. This will complete the construction of λ as a probability measure (countably additive, that is) on $\mathscr{B}$, and the construction is fundamental to all that follows. For example, the set N of normal numbers lies in $\mathscr{B}$ (Example 2.6), and it will turn out that $\lambda(N) = 1$, as probabilistic intuition requires. (In Chapter 2, λ will be defined for sets outside the unit interval as well.)

It is well to pause here and consider just what is involved in the construction of Lebesgue measure on the Borel sets of the unit interval. That length defines a finitely additive set function on the class $\mathscr{I}$ of intervals in (0, 1] is a consequence of Theorem 1.3 for the case of only finitely many intervals and thus involves only the most elementary properties of the real number system. But proving countable additivity on $\mathscr{I}$ requires the deeper property of compactness (the Heine-Borel theorem). Once λ has been proved countably additive on $\mathscr{I}$, extending it to $\mathscr{B}_0$ by the definition (2.12) presents no real difficulty: the arguments involving (2.13) and (2.14) are easy. Difficulties again arise, however, in the further extension of λ from $\mathscr{B}_0$ to $\mathscr{B} = \sigma(\mathscr{B}_0)$, and here new ideas are again required. These ideas are the subject of Section 3, where it is shown that any probability measure on any field can be extended to the generated σ-field.

Sequence Space†

Let S be a finite set of points regarded as the possible outcomes of a simple observation or experiment. For tossing a coin, S can be $\{\mathrm{H}, \mathrm{T}\}$ or $\{0, 1\}$; for rolling a die, $S = \{1, \ldots, 6\}$; in information theory, S plays the role of a finite alphabet. Let $\Omega = S^{\infty}$ be the space of all infinite sequences

$$\omega = (z_1(\omega), z_2(\omega), \ldots) \tag{2.15}$$

†The ideas that follow are basic to probability theory and are used further on, in particular in Section 24 and (in more elaborate form) Section 36. On a first reading, however, one might prefer to skip to Section 3 and return to this topic as the need arises.

of elements of S: $z_k(\omega) \in S$ for all $\omega \in S^\infty$ and $k \geq 1$. The sequence (2.15) can be viewed as the result of repeating infinitely often the simple experiment represented by S. For $S = \{0, 1\}$, the space S^∞ is closely related to the unit interval; compare (1.8) and (2.15).

The space S^∞ is an infinite-dimensional Cartesian product. Each $z_k(\cdot)$ is a mapping of S^∞ onto S; these are the *coordinate functions*, or the *natural projections*. Let $S^n = S \times \cdots \times S$ be the Cartesian product of n copies of S; it consists of the n-long sequences $(u_1, \ldots, u_n)$ of elements of S. For such a sequence, the set

$$[\omega\colon (z_1(\omega), \ldots, z_n(\omega)) = (u_1, \ldots, u_n)] \tag{2.16}$$

represents the event that the first n repetitions of the experiment give the outcomes $u_1, \ldots, u_n$ in sequence. A *cylinder of rank n* is a set of the form

$$A = [\omega\colon (z_1(\omega), \ldots, z_n(\omega)) \in H], \tag{2.17}$$

where $H \subset S^n$. Note that A is nonempty if H is. If H is a singleton in S^n, (2.17) reduces to (2.16), which can be called a *thin* cylinder.

Let $\mathscr{C}_0$ be the class of cylinders of all ranks. Then $\mathscr{C}_0$ *is a field*: S^∞ and the empty set have the form (2.17) for $H = S^n$ and for $H = \emptyset$. If H is replaced by $S^n - H$, then (2.17) goes into its complement, and hence $\mathscr{C}_0$ is closed under complementation. As for unions, consider (2.17) together with

$$B = [\omega\colon (z_1(\omega), \ldots, z_m(\omega)) \in I], \tag{2.18}$$

a cylinder of rank m. Suppose that $n \leq m$ (symmetry); if H' consists of the sequences $(u_1, \ldots, u_m)$ in S^m for which the truncated sequence $(u_1, \ldots, u_n)$ lies in H, then (2.17) has the alternative form

$$A = [\omega\colon (z_1(\omega), \ldots, z_m(\omega)) \in H']. \tag{2.19}$$

Since it is now clear that

$$A \cup B = [\omega\colon (z_1(\omega), \ldots, z_m(\omega)) \in H' \cup I] \tag{2.20}$$

is also a cylinder, $\mathscr{C}_0$ is closed under the formation of finite unions and hence is indeed a field.

Let $p_u, u \in S$, be probabilities on S—nonnegative and summing to 1. Define a set function P on $\mathscr{C}_0$ (it will turn out to be a probability measure) in this way: For a cylinder A given by (2.17), take

$$P(A) = \sum_H p_{u_1} \cdots p_{u_n}, \tag{2.21}$$

the sum extending over all the sequences $(u_1, \ldots, u_n)$ in H. As a special case,

$$P[\omega\colon (z_1(\omega), \ldots, z_n(\omega)) = (u_1, \ldots, u_n)] = p_{u_1} \cdots p_{u_n}. \tag{2.22}$$

Because of the products on the right in (2.21) and (2.22), P is called *product measure*; it provides a model for an infinite sequence of independent repetitions of the simple experiment represented by the probabilities p_u on S. In the case where $S = \{0, 1\}$ and $p_0 = p_1 = \frac{1}{2}$, it is a model for independent tosses of a fair coin, an alternative to the model used in Section 1.

The definition (2.21) presents a consistency problem, since the cylinder A will have other representations. Suppose that A is also given by (2.19). If $n = m$, then H and H' must coincide, and there is nothing to prove. Suppose then (symmetry) that $n < m$. Then H' must consist of those $(u_1, \ldots, u_m)$ in S^m for which $(u_1, \ldots, u_n)$ lies in H: $H' = H \times S^{m-n}$. But then

$$\begin{aligned} \sum_{H'} p_{u_1} \cdots p_{u_n} p_{u_{n+1}} \cdots p_{u_m} &= \sum_{H} p_{u_1} \cdots p_{u_n} \sum_{S^{m-n}} p_{u_{n+1}} \cdots p_{u_m} \\ &= \sum_{H} p_{u_1} \cdots p_{u_n}. \end{aligned} \tag{2.23}$$

The definition (2.21) is therefore consistent. And finite additivity is now easy: Suppose that A and B are disjoint cylinders given by (2.17) and (2.18). Suppose that $n \le m$, and put A in the form (2.19). Since A and B are disjoint, H' and I must be disjoint as well, and by (2.20),

$$P(A \cup B) = \sum_{H' \cup I} p_{u_1} \cdots p_{u_m} = P(A) + P(B). \tag{2.24}$$

Taking $H = S^n$ in (2.21) shows that $P(S^\infty) = 1$. Therefore, (2.21) *defines a finitely additive probability measure on the field* $\mathscr{C}_0$.

Now, P is countably additive on $\mathscr{C}_0$, but this requires no further argument, because of the following completely general result.

THEOREM 2.3

Every finitely additive probability measure on the field $\mathscr{C}_0$ of cylinders in S^∞ is in fact countably additive.

The proof depends on this fundamental fact:

Lemma. *If $A_n \downarrow A$, where the A_n are nonempty cylinders, then A is nonempty.*

Proof of Theorem 2.3. Assume that the lemma is true, and apply Example 2.10 to the measure P in question: If $A_n \downarrow \varnothing$ for sets in $\mathscr{C}_0$ (cylinders) but $P(A_n)$

does *not* converge to 0, then $P(A_n) \geq \epsilon > 0$ for some ϵ. But then the A_n are nonempty, which by the lemma makes $A_n \downarrow \emptyset$ impossible. ■

Proof of the Lemma.† Suppose that A_t is a cylinder of rank m_t, say

$$A_t = [\omega\colon (z_1(\omega), \ldots, z_{m_t}(\omega)) \in H_t], \tag{2.25}$$

where $H_t \subset S^{m_t}$. Choose a point ω_n in A_n, which is nonempty by assumption. Write the components of the sequences in a square array:

$$\begin{array}{cccc} z_1(\omega_1) & z_1(\omega_2) & z_1(\omega_3) & \cdots \\ z_2(\omega_1) & z_2(\omega_2) & z_2(\omega_3) & \cdots \\ \vdots & \vdots & \vdots & \end{array} \tag{2.26}$$

The nth *column* of the array gives the components of ω_n.

Now argue by a modification of the diagonal method [A14]. Since S is finite, some element u_1 of S appears infinitely often in the first row of (2.26): for an increasing sequence $\{n_{1,k}\}$ of integers, $z_1(\omega_{n_{1,k}}) = u_1$ for all k. By the same reasoning, there exist an increasing subsequence $\{n_{2,k}\}$ of $\{n_{1,k}\}$ and an element u_2 of S such that $z_2(\omega_{n_{2,k}}) = u_2$ for all k. Continue. If $n_k = n_{k,k}$, then $z_r(\omega_{n_k})u_r$ for $k \geq r$, and hence $(z_1(\omega_{n_k}), \ldots, z_r(\omega_{n_k})) = (u_1, \ldots, u_r)$ for $k \geq r$.

Let ω° be the element of S^∞ with components u_r: $\omega^\circ = (u_1, u_2, \ldots) = (z_1(\omega^\circ), z_2(\omega^\circ), \ldots)$. Let t be arbitrary. If $k \geq t$, then (n_k is increasing) $n_k \geq t$ and hence $\omega_{n_k} \in A_{n_k} \subset A_t$. It follows by (2.25) that, for $k \geq t, H_t$ contains the point $(z_1(\omega_{n_k}), \ldots, z_{m_t}(\omega_{n_k}))$ of S^{m_t}. But for $k \geq m_t$, this point is identical with $(z_1(\omega^\circ), \ldots, z_{m_t}(\omega^\circ))$, which therefore lies in H_t. Thus ω° is a point common to all the A_t. ■

Let $\mathscr{C}$ be the σ-field in S^∞ generated by $\mathscr{C}_0$. By the general theory of the next section, the probability measure P defined on $\mathscr{C}_0$ by (2.21) extends to $\mathscr{C}$. The term *product measure*, properly speaking, applies to the extended P. Thus $(S^\infty, \mathscr{C}, P)$ is a probability space, one important in ergodic theory (Section 24).

Suppose that $S = \{0, 1\}$ and $p_0 = p_1 = \frac{1}{2}$. In this case, $(S^\infty, \mathscr{C}, P)$ is closely related to $((0, 1], \mathscr{B}, \lambda)$, although there are essential differences. The sequence (2.15) can end in 0's, but (1.8) cannot. Thin cylinders are like dyadic intervals, but the sets in $\mathscr{C}_0$ (the cylinders) correspond to the finite disjoint unions of intervals with dyadic endpoints, a field somewhat smaller than $\mathscr{B}_0$. While nonempty sets in $\mathscr{B}_0$ (for example, $\left(\frac{1}{2}, \frac{1}{2} + 2^{-n}\right]$) can contract to the empty set, nonempty sets in $\mathscr{C}_0$ cannot. The lemma above plays here the role the Heine-Borel theorem plays in the proof of Theorem 1.3. The product probability measure constructed here on $\mathscr{C}_0$ (in the case $S = \{0, 1\}, p_0 = p_1 = \frac{1}{2}$, that is) is analogous to Lebesgue

†The lemma is a special case of Tychonov's theorem: If S is given the discrete topology, the topological product S^∞ is compact (and the cylinders are closed).

measure on $\mathscr{B}_0$. But a finitely additive probability measure on $\mathscr{B}_0$ can fail to be countably additive,[†] which cannot happen in $\mathscr{C}_0$.

Constructing σ-Fields[‡]

The σ-field $\sigma(\mathscr{A})$ generated by $\mathscr{A}$ was defined from above or from the outside, so to speak, by intersecting all the σ-fields that contain $\mathscr{A}$ (including the σ-field consisting of all the subsets of Ω). Can $\sigma(\mathscr{A})$ somehow be constructed from the inside by repeated finite and countable set-theoretic operations starting with sets in $\mathscr{A}$?

For any class $\mathscr{H}$ of sets in Ω let $\mathscr{H}^*$ consist of the sets in $\mathscr{H}$, the complements of sets in $\mathscr{H}$, and the finite and countable unions of sets in $\mathscr{H}$. Given a class $\mathscr{A}$, put $\mathscr{A}_0 = \mathscr{A}$ and define $\mathscr{A}_1, \mathscr{A}_2, \ldots$ inductively by

$$\mathscr{A}_n = \mathscr{A}_{n-1}^*. \tag{2.27}$$

That each $\mathscr{A}_n$ is contained in $\sigma(\mathscr{A})$ follows by induction. One might hope that $\mathscr{A}_n = \sigma(\mathscr{A})$ for some n, or at least that $\cup_{n=0}^{\infty} \mathscr{A}_n = \sigma(\mathscr{A})$. But this process applied to the class of intervals fails to account for all the Borel sets.

Let $\mathscr{I}_0$ consist of the empty set and the intervals in $\Omega = (0, 1]$ with rational endpoints, and define $\mathscr{I}_n = \mathscr{I}_{n-1}^*$ for $n = 1, 2, \ldots$. *It will be shown that* $\bigcup_{n=0}^{\infty} \mathscr{I}_n$ *is strictly smaller than* $\mathscr{B} = \sigma(\mathscr{I}_0)$.

If a_n and b_n are rationals decreasing to a and b, then $(a, b] = \bigcup_m \bigcap_n (a_m, b_n] = \bigcup_m (\bigcup_n (a_m, b_n]^c)^c \in \mathscr{I}_4$. The result would therefore not be changed by including in $\mathscr{I}_0$ *all* the intervals in (0.1].

To prove $\bigcup_{n-0}^{\infty} \mathscr{I}_n$ smaller than $\mathscr{B}$, first put

$$\psi(A_1, A_2, \ldots) = A_1^c \cup A_2 \cup A_3 \cup A_4 \cup \cdots. \tag{2.28}$$

Since $\mathscr{I}_{n-1}$ contains $\Omega = (0, 1]$ and the empty set, every element of $\mathscr{I}_n$ has the form (2.28) for some sequence $A_1, A_2, \ldots$ of sets in $\mathscr{I}_{n-1}$. Let every positive integer appear exactly once in the square array

$$\begin{matrix} m_{11} & m_{12} & \cdots \\ m_{21} & m_{22} & \cdots \\ \vdots & \vdots & \end{matrix}$$

Inductively define

$$\Phi_0(A_1, A_2, \ldots) = A_1, \tag{2.29}$$

$$\Phi_n(A_1, A_2 \ldots) = \psi(\Phi_{n-1}(A_{m_{11}}, A_{m_{12}}, \ldots), \Phi_{n-1}(A_{m_{21}}, A_{m_{22}} \ldots), \ldots), \quad n = 1, 2, \ldots.$$

[†]See Problem 2.15.

[‡]This topic may be omitted.

It follows by induction that every element of $\mathscr{I}_n$ has the form $\Phi_n(A_1, A_2, \ldots)$ for some sequence of sets in $\mathscr{I}_0$. Finally, put

$$\Phi(A_1, A_2, \ldots) = \Phi_1(A_{m_{11}}, A_{m_{12}}, \ldots) \cup \Phi_2(A_{m_{21}}, A_{m_{22}}, \ldots) \cup \cdots. \tag{2.30}$$

Then every element of $\bigcup_{n=0}^{\infty} \mathscr{I}_n$ has the form (2.30) for some sequence $A_1, A_2, \ldots$ of sets in $\mathscr{I}_0$.

If $A_1, A_2, \ldots$ are in $\mathscr{B}$, then (2.28) is in $\mathscr{B}$; it follows by induction that each $\Phi_n(A_1, A_2, \ldots)$ is in $\mathscr{B}$ and therefore that (2.30) is in $\mathscr{B}$.

With each ω in (0, 1] associate the sequence $(\omega_1, \omega_2, \ldots)$ of positive integers such that $\omega_1 + \cdots + \omega_k$ is the position of the kth 1 in the nonterminating dyadic expansion of ω (the smallest n for which $\sum_{i=1}^{n} d_j(\omega) = k$). Then $\omega \leftrightarrow (\omega_1, \omega_2, \ldots)$ is a one-to-one correspondence between (0, 1] and the set of all sequences of positive integers. Let $I_1, I_2, \ldots$ be an enumeration of the sets in $\mathscr{I}_0$, put $\varphi(\omega) = \Phi(I_{\omega_1}, I_{\omega_2}, \ldots)$, and define $B = [\omega\colon\ \omega \notin \varphi(\omega)]$. It will be shown that B is a Borel set but is not contained in any of the $\mathscr{I}_n$.

Since ω lies in B if and only if ω lies outside $\varphi(\omega)$, $B \neq \varphi(\omega)$ for every ω. But every element of $\bigcup_{n=0}^{\infty} \mathscr{I}_n$ has the form (2.30) for some sequence in $\mathscr{I}_0$ and hence has the form $\varphi(\omega)$ for some ω. Therefore, B is not a member of $\bigcup_{n=0}^{\infty} \mathscr{I}_1$.

It remains to show that B is a Borel set. Let $D_k = [\omega\colon\ \omega \in I_{\omega_k}]$. Since $L_k(n) = [\omega\colon\ \omega_1 + \cdots + \omega_k = n] = [\omega\colon\ \sum_{j=1}^{n-1} d_j(\omega) < k = \sum_{j=1}^{n} d_j(\omega)]$ is a Borel set, so are $[\omega\colon\ \omega_k = n] = \bigcup_{m=1}^{\infty} L_{k-1}(m) \cap L_k(m+n)$ and

$$D_k = [\omega\colon\ \omega \in I_{w_k}] = \bigcup_n ([\omega\colon\ \omega_k = n] \cap I_n).$$

Suppose that it is shown that

$$[\omega\colon\ \omega \in \Phi_n(I_{\omega_{u_1}}, I_{\omega_{u_2}} \ldots)] = \Phi_n(D_{u_1}, D_{u_{21}}, \ldots) \tag{2.31}$$

for every n and every sequence $u_1, u_2, \ldots$ of positive integers. It will then follow from the definition (2.30) that

$$B^c = [\omega\colon\ \omega \in \varphi(\omega)] = \bigcup_{n-1}^{\infty} \left[\omega\colon\ \omega \in \Phi_n(I_{\omega_{m_{n1}}}, I_{\omega_{m_{n2}}}, \ldots)\right]$$

$$= \bigcup_{n=1}^{\infty} \Phi_n(D_{m_{n1}}, D_{m_{n2}}, \ldots) = \Phi(D_1, D_2, \ldots).$$

But as remarked above, (2.30), is a Borel set if the A_n are. Therefore, (2.31) will imply that B^c and B are Borel sets.

If $n = 0$, (2.31) holds because it reduces by (2.29) to $[\omega\colon\ \omega \in I_{\omega_{u_1}}]I = D_{u_1}$. Suppose that (2.31) holds with $n-1$ in place of n. Consider the condition

$$\omega \in \Phi_{n-1}(I_{\omega_{um_{k1}}}, I_{\omega_{um_{k2}}}, \ldots). \tag{2.32}$$

By (2.28) and (2.29), a necessary and sufficient condition for $\omega \in \Phi_n(I_{\omega_{u_1}}, I_{\omega_{u_2}}, \ldots)$ is that either (2.32) is false for $k = 1$ or else (2.32) is true for some k exceeding 1. But by the induction hypothesis, (2.32) and its negation can be replaced by $\omega \in \Phi_{n-1}(D_{u_{m_{k1}}}, D_{u_{m_{k2}}}, \ldots)$ and its negation. Therefore, $\omega \in \Phi_n(I_{\omega_{u1}}, I_{\omega_{u2}}, \ldots)$ if and only if $\omega \in \Phi_n(D_{u_1}, D_{u_2}, \ldots)$.

Thus $\bigcup_n \mathscr{I}_n \neq \mathscr{B}$, and there are Borel sets that cannot be arrived at from the intervals by any finite sequence of set-theoretic operations, each operation being finite or countable. It can even be shown that there are Borel sets that cannot be arrived at by any *countable* sequence of these operations. On the other hand, every Borel set can be arrived at by a countable *ordered set* of these operations if it is not required that they be performed in a simple *sequence*. The proof of this statement—and indeed even a precise explanation of its meaning—depends on the theory of infinite ordinal numbers.[†]

PROBLEMS

2.1. Define $x \vee y = \max\{x, y\}$, and for a collection $\{x_\alpha\}$ define $\vee_\alpha x_\alpha = \sup_\alpha x_\alpha$; define $x \wedge y = \min\{x, y\}$ and $\wedge_\alpha x_\alpha = \inf_\alpha x_\alpha$. Prove that $I_{A\cup B} = I_A \vee I_B, I_{A\cap B} = 1_A \wedge I_B, I_{A^c} = 1 - I_A$, and $I_{A\Delta B} = |I_A - I_B|$ in the sense that there is equality at each point of Ω. Show that $A \subset B$ if and only if $I_A \leq I_B$ pointwise. Check the equation $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ and deduce the distribute law

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. By similar arguments prove that

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

$$A\Delta C \subset (A\Delta B) \cup (B\Delta C),$$

$$\left(\bigcup_n A_n\right)^c = \bigcap_n A_n^c,$$

$$\left(\bigcap_n A_n\right)^c = \bigcup_n A_n^c.$$

2.2. Let $A_1, \ldots, A_n$ be arbitrary events, and put $U_k = \bigcup(A_{i_1} \cap \cdots \cap A_{i_k})$ and $I_k = \bigcap(A_{i_1}, \bigcup \cdots \bigcup A_{i_k})$, where the union and intersection extend over all the k-tuples satisfying $1 \leq l_1 < \cdots < i_k \leq n$. Show that $U_k = I_{n-k+1}$.

2.3. **(a)** Suppose that $\Omega \in \mathscr{F}$ and that $A, B \in \mathscr{F}$ implies $A - B = A \cap B^c \in \mathscr{F}$. Show that $\mathscr{F}$ is a field.

[†]See Problem 2.22.

(b) Suppose that $\Omega \in \mathscr{F}$ and that $\mathscr{F}$ is closed under the formation of complements and finite *disjoint* unions. Show that $\mathscr{F}$ need not be a field.

2.4. Let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be classes of sets in a common space Ω.

(a) Suppose that $\mathscr{F}_n$ are fields satisfying $\mathscr{F}_n \subset \mathscr{F}_{n+1}$. Show that $\bigcup_{n=1}^{\infty} \mathscr{F}_n$ is a field.

(b) Suppose that $\mathscr{F}_n$ are σ-fields satisfying $\mathscr{F}_n \subset \mathscr{F}_{n+1}$. Show by example that $\bigcup_{n=1}^{\infty} \mathscr{F}_n$ need not be a σ-field.

2.5. The field $f(\mathscr{A})$ generated by a class $\mathscr{A}$ in Ω is defined as the intersection of all fields in Ω containing $\mathscr{A}$.

(a) Show that $f(\mathscr{A})$ is indeed a field, that $\mathscr{A} \subset f(\mathscr{A})$, and that $f(\mathscr{A})$ is minimal in the sense that if $\mathscr{G}$ is a field and $\mathscr{A} \subset \mathscr{G}$, then $f(\mathscr{A}) \subset \mathscr{G}$.

(b) Show that for nonempty $\mathscr{A}$, $f(\mathscr{A})$ is the class of sets of the form $\bigcup_{i=1}^{m} \bigcap_{j=1}^{n_i} A_{ij}$, where for each i and j either $A_{ij} \in \mathscr{A}$ or $A_{ij}^c \in \mathscr{A}$, and where the m sets $\bigcap_{j=1}^{n_i} A_{ij}, 1 \le i \le m$, are disjoint. The sets in $f(\mathscr{A})$ can thus be explicitly presented, which is not in general true of the sets in $\sigma(\mathscr{A})$.

2.6. $\uparrow$

(a) Show that if $\mathscr{A}$ consists of the singletons, then $f(\mathscr{A})$ is the field in Example 2.3.

(b) Show that $f(\mathscr{A}) \subset \sigma(\mathscr{A})$, that $f(\mathscr{A}) = \sigma(\mathscr{A})$ if $\mathscr{A}$ is finite, and that $\sigma(f(\mathscr{A})) = \sigma(\mathscr{A})$.

(c) Show that if $\mathscr{A}$ is countable, then $f(\mathscr{A})$ is countable.

(d) Show for fields $\mathscr{F}_1$ and $\mathscr{F}_2$ that $f(\mathscr{F}_1 \cup \mathscr{F}_2)$ consists of the finite disjoint unions of sets $A_1 \cap A_2$ with $A_i \in \mathscr{F}_i$. Extend.

2.7. 2.5 $\uparrow$ Let H be a a set lying outside $\mathscr{F}$, where $\mathscr{F}$ is a field [or σ-field]. Show that the field [or σ-field] generated by $\mathscr{F} \cup \{H\}$ consists of sets of the form

$$(H \cap A) \cup (H^c \cap B), \quad A, B \in \mathscr{F}. \tag{2.33}$$

2.8. Suppose for each A in $\mathscr{A}$ that A^c is a countable union of elements of $\mathscr{A}$. The class of intervals in (0, 1] has this property. Show that $\sigma(\mathscr{A})$ coincides with the smallest class over $\mathscr{A}$ that is closed under the formation of countable unions and intersections.

2.9. Show that, if $B \in \sigma(\mathscr{A})$, then there exists a countable subclass $\mathscr{A}_B$ of $\mathscr{A}$ such that $B \in \sigma(\mathscr{A}_B)$.

2.10. **(a)** Show that if $\sigma(\mathscr{A})$ contains every subset of Ω, then for each pair ω and ω' of distinct points in Ω there is in $\mathscr{A}$ an A such that $I_A(\omega) \ne I_A(\omega')$.

(b) Show that the reverse implication holds if Ω is countable.

(c) Show by example that the reverse implication need not hold for uncountable Ω.

2.11. A σ-field is *countably generated*, or *separable*, if it is generated by some countable class of sets.

(a) Show that the σ-field $\mathscr{B}$ of Borel sets is countably generated.

(b) Show that the σ-field of Example 2.4 is countably generated if and only if Ω is countable.

(c) Suppose that $\mathscr{F}_1$ and $\mathscr{F}_2$ are σ-fields, $\mathscr{F}_1 \subset \mathscr{F}_2$, and $\mathscr{F}_2$ is countably generated. Show by example that $\mathscr{F}_1$ may not be countably generated.

2.12. Show that a σ-field cannot be countably infinite—its cardinality must be finite or else at least that of the continuum. Show by example that a field can be countably infinite.

2.13. **(a)** Let $\mathscr{F}$ be the field consisting of the finite and the cofinite sets in an infinite Ω, and define P on $\mathscr{F}$ by taking $P(A)$ to be 0 or 1 as A is finite or cofinite. (Note that P is not well defined if Ω is finite.) Show that P is finitely additive.

(b) Show that this P is not countably additive if Ω is countably infinite.

(c) Show that this P is countably additive if Ω is uncountable.

(d) Now let $\mathscr{F}$ be the σ-field consisting of the countable and the cocountable sets in an uncountable Ω, and define P on $\mathscr{F}$ by taking $P(A)$ to be 0 or 1 as A is countable or cocountable. (Note that P is not well defined if Ω is countable.) Show that P is countably additive.

2.14. In (0, 1] let $\mathscr{F}$ be the class of sets that either (i) are of the first category [A15] or (ii) have complement of the first category. Show that $\mathscr{F}$ is a σ-field. For A in $\mathscr{F}$, take $P(A)$ to be 0 in case (i) and 1 in case (ii). Show that P is countably additive.

2.15. On the field $\mathscr{B}_0$ in (0, 1] define $P(A)$ to be 1 or 0 according as there does or does not exist some positive ϵ_A (depending on A) such that A contains the interval $(\frac{1}{2}, \frac{1}{2} + \epsilon_A]$. Show that P is finitely but not countably additive. No such example is possible for the field $\mathscr{C}_0$ in S^∞ (Theorem 2.3).

2.16. **(a)** Suppose that P is a probability measure on a field $\mathscr{F}$. Suppose that $A_t \in \mathscr{F}$ for $t > 0$, that $A_s \subset A_t$ for $s < t$, and that $A = \bigcup_{t>0} A_t \in \mathscr{F}$. Extend Theorem 2.1(i) by showing that $P(A_t) \uparrow P(A)$ as $t \to \infty$. Show that A necessarily lies in $\mathscr{F}$ if it is a σ-field.

(b) Extend Theorem 2.1(ii) in the same way.

2.17. Suppose that P is a probability measure on a field $\mathscr{F}$, that $A_1, A_2, \ldots$, and $A = \bigcup_n A_n$ lie in $\mathscr{F}$, and that the A_n are nearly disjoint in the sense that $P(A_m \cap A_n) = 0$ for $m \neq n$. Show that $P(A) = \sum_n P(A_n)$.

2.18. *Stochastic arithmetic*. Define a set function P_n on the class of all subsets of $\Omega = \{1, 2, \ldots\}$ by

$$P_n(A) = \frac{1}{n}\#[m\colon 1 \le m \le n, m \in A]; \tag{2.34}$$

among the first n integers, the proportion that lie in A is just $P_n(A)$. Then P_n is a discrete probability measure. The set A has *density*

$$D(A) = \lim_n P_n(A), \tag{2.35}$$

provided this limit exists. Let $\mathscr{D}$ be the class of sets having density.

(a) Show that D is finitely but not countably additive on $\mathscr{D}$.

(b) Show that $\mathscr{D}$ contains the empty set and Ω and is closed under the formation of complements, proper differences, and finite disjoint unions, but is not closed under the formation of countable disjoint unions or of finite unions that are not disjoint.

(c) Let $\mathscr{M}$ consist of the periodic sets $M_a = [ka\colon k = 1, 2, \ldots]$. Observe that

$$P_n(M_a) = \frac{1}{n}\left\lfloor \frac{n}{a} \right\rfloor \to \frac{1}{a} = D(M_a). \tag{2.36}$$

Show that the field $f(\mathscr{M})$ generated by $\mathscr{M}$ (see Problem 2.5) is contained in $\mathscr{D}$. Show that D is completely determined on $f(\mathscr{M})$ by the value it gives for each a to the event that m is divisible by a.

(d) Assume that $\sum p^{-1}$ diverges (sum over all primes; see Problem 5.20(e)) and prove that D, although finitely additive, is not countably additive on the field $f(M)$.

(e) Euler's function $\varphi(n)$ is the number of positive integers less than n and relatively prime to it. Let $p_1, \ldots, p_r$ be the distinct prime factors of n; from the inclusion-exclusion formula for the events $[m\colon p_i | m]$, (2.36), and the fact that the p_i divide n, deduce

$$\frac{\varphi(n)}{n} = \prod_{p|n}\left(1 - \frac{1}{p}\right). \tag{2.37}$$

(f) Show for $0 \le x \le 1$ that $D(A) = x$ for some A.

(g) Show that D is translation invariant: If $B = [m + 1\colon m \in A]$, then B has a density if and only if A does, in which case $D(A) = D(B)$.

2.19. A probability measure space $(\Omega, \mathscr{F}, P)$ is *nonatomic* if $P(A) > 0$ implies that there exists a B such that $B \subset A$ and $0 < P(B) < P(A)$ (A and B in $\mathscr{F}$, of course).

(a) Assuming the existence of Lebesgue measure λ on $\mathscr{B}$, prove that it is nonatomic.

(b) Show in the nonatomic case that $P(A) > 0$ and $\epsilon > 0$ imply that there exists a B such that $B \subset A$ and $0 < P(B) < \epsilon$.

(c) Show in the nonatomic case that $0 \le x \le P(A)$ implies that there exists a B such that $B \subset A$ and $P(B) = x$. *Hint*: Inductively define classes $\mathscr{H}_n$, numbers h_n, and sets H_n by $\mathscr{H}_0 = \{\emptyset\} = \{H_0\}$, $\mathscr{H}_n = [H: H \subset A - \bigcup_{k<n} H_k, P(\bigcup_{k<n} H_k) + P(H) \le x]$, $h_n = \sup[P(H): H \in \mathscr{H}_n]$, and $P(H_n) > h_n - n^{-1}$. Consider $\bigcup_k H_k$.

(d) Show in the nonatomic case that, if $p_1, p_2, \ldots$ are nonnegative and add to 1, then A can be decomposed into sets $B_1, B_2, \ldots$ such that $P(B_n) = p_n P(A)$.

2.20. Generalize the construction of product measure: For $n = 1, 2, \ldots$, let S_n be a finite space with given probabilities $p_{nu}, u \in S_n$. Let $S_1 \times S_2 \times \cdots$ be the space of sequences (2.15), where now $z_k(\omega) \in S_k$. Define P on the class of cylinders, appropriately defined, by using the product $p_{1u_1} \cdots p_{nu_n}$ on the right in (2.21). Prove P countably additive on $\mathscr{C}_0$, and extend Theorem 2.3 and its lemma to this more general setting. Show that the lemma fails if any of the S_n are infinite.

2.21. **(a)** Suppose that $\mathscr{A} = \{A_1, A_2, \ldots\}$ is a countable partition of Ω. Show (see (2.27)) that $\mathscr{A}_1 = \mathscr{A}_0^* = \mathscr{A}^*$ coincides with $\sigma(\mathscr{A})$. This is a case where $\sigma(\mathscr{A})$ can be constructed "from the inside."

(b) Show that the set of normal numbers lies in $\mathscr{I}_6$.

(c) Show that $\mathscr{H}^* = \mathscr{H}$ if and only if $\mathscr{H}$ is a σ-field. Show that $\mathscr{I}_{n-1}$ is strictly smaller than $\mathscr{I}_n$ for all n.

2.22. Extend (2.27) to infinite ordinals α by defining $\mathscr{A}_\alpha = (\bigcup_{\beta<\alpha} \mathscr{A}_\beta)^*$. Show that, if Ω is the first uncountable ordinal, then $\bigcup_{\alpha<\Omega} \mathscr{A}_\alpha = \sigma(\mathscr{A})$. Show that, if the cardinality of $\mathscr{A}$ does not exceed that of the continuum, then the same is true of $\sigma(\mathscr{A})$. Thus $\mathscr{B}$ has the power of the continuum.

2.23. $\uparrow$ Extend (2.29) to ordinals $\alpha < \Omega$ as follows. Replace the right side of (2.28) by $\bigcup_{n=1}^{\infty}(A_{2n-1} \cup A_{2n}^c)$. Suppose that Φ_β is defined for $\beta < \alpha$. Let $\beta_\alpha(1), \beta_\alpha(2), \ldots$ be a sequence of ordinals such that $\beta_\alpha(n) < \alpha$ and such that if $\beta < \alpha$, then $\beta = \beta_\alpha(n)$ for infinitely many even n and for infinitely many odd n; define

$$\Phi_\alpha(A_1, A_2, \ldots) = \psi(\Phi_{\beta_\alpha(1)}(A_{m_{11}}, A_{m_{12}}, \ldots), \Phi_{\beta_\alpha(2)}(A_{m_{21}}, A_{m_{22}}, \ldots), \ldots). \tag{2.38}$$

Prove by transfinite induction that (2.38) is in $\mathscr{B}$ if the A_n are, that every element of $\mathscr{I}_\alpha$ has the form (2.38) for sets A_n in $\mathscr{I}_0$, and that (2.31) holds with α in place of n. Define $\varphi_\alpha(\omega) = \Phi_\alpha(I_{\omega_1}, I_{\omega_2}, \ldots)$, and show that $B_\alpha = [\omega: \omega \notin \varphi_\alpha(\omega)]$ lies in $\mathscr{B} - \mathscr{I}_\alpha$ for $\alpha < \Omega$. Show that $\mathscr{I}_\alpha$ is strictly smaller than $\mathscr{I}_\beta$ for $\alpha < \beta \le \Omega$.

SECTION 3 EXISTENCE AND EXTENSION

The main theorem to be proved here may be compactly stated this way:

THEOREM 3.1
A probability measure on a field has a unique extension to the generated σ-field.

In more detail the assertion is this: Suppose that P is a probability measure on a field $\mathscr{F}_0$ of subsets of Ω, and put $\mathscr{F} = \sigma(\mathscr{F}_0)$. Then there exists a probability measure Q on $\mathscr{F}$ such that $Q(A) = P(A)$ for $A \in \mathscr{F}_0$. Further, if Q' is another probability measure on $\mathscr{F}$ such that $Q'(A) = P(A)$ for $A \in \mathscr{F}_0$, then $Q'(A) = Q(A)$ for $A \in \mathscr{F}$.

Although the measure extended to $\mathscr{F}$ is usually denoted by the same letter as the original measure on $\mathscr{F}_0$, they are really different set functions, since they have different domains of definition. The class $\mathscr{F}_0$ is only assumed finitely additive in the theorem, but the set function P on it must be assumed countably additive (since this of course follows from the conclusion of the theorem).

As shown in Theorem 2.2, λ (initially defined for intervals as length: $\lambda(I) = |I|$) extends to a probability measure on the field $\mathscr{B}_0$ of finite disjoint unions of subintervals of (0, 1]. By Theorem 3.1, λ extends in a unique way from $\mathscr{B}_0$ to $\mathscr{B} = \sigma(\mathscr{B}_0)$, the class of Borel sets in (0,1]. The extended λ is *Lebesgue measure* on the unit interval. Theorem 3.1 has many other applications as well.

The uniqueness in Theorem 3.1 will be proved later; see Theorem 3.3. The first project is to prove that an extension does exist.

Construction of the Extension

Let P be a probability measure on a field $\mathscr{F}_0$. The construction following extends P to a class that in general is much larger than $\sigma(\mathscr{F}_0)$ but nonetheless does not in general contain all the subsets of Ω.

For each subset A of Ω, define its *outer measure* by

$$P^*(A) = \inf \sum_n P(A_n), \tag{3.1}$$

where the infimum extends over all finite and infinite sequences $A_1, A_2, \ldots$ of $\mathscr{F}_0$-sets satisfying $A \subset \cup_n A_n$. If the A_n form an efficient covering of A, in the sense that they do not overlap one another very much or extend much beyond A, then $\sum_n P(A_n)$ should be a good outer approximation to the measure of A if A is indeed to have a measure assigned it at all. Thus (3.1) represents a first attempt to assign a measure to A.

Because of the rule $P(A^c) = 1 - P(A)$ for complements (see (2.6)), it is natural in approximating A from the inside to approximate the complement A^c

from the outside instead and then subtract from 1:

$$P_*(A) = 1 - P^*(A^c). \tag{3.2}$$

This, the *inner measure* of A, is a second candidate for the measure of A.[†] A plausible procedure is to assign measure to those A for which (3.1) and (3.2) agree, and to take the common value $P^*(A) = P_*(A)$ as the measure. Since (3.1) and (3.2) agree if and only if

$$P^*(A) + P^*(A^c) = 1, \tag{3.3}$$

the procedure would be to consider the class of A satisfying (3.3) and use $P^*(A)$ as the measure.

It turns out to be simpler to impose on A the more stringent requirement that

$$P^*(A \cap E) - P^*(A^c \cap E) = P^*(E) \tag{3.4}$$

hold for every set E; (3.3) is the special case $E = \Omega$, because it will turn out that $P^*(\Omega) = 1$.[‡] A set A is called *P^*-measurable* if (3.4) holds for all E; let $\mathcal{M}$ be the class of such sets. What will be shown is that $\mathcal{M}$ contains $\sigma(\mathcal{T}_0)$ and that the restriction of P^* to $\sigma(\mathcal{T}_0)$ is the required extension of P.

The set function P^* has four properties that will be needed:

(i) $P^*(\emptyset) = 0$;

(ii) P^* is nonnegative: $P^*(A) \geq 0$ for every $A \subset \Omega$;

(iii) P^* is monotone: $A \subset B$ implies $P^*(A) \leq P^*(B)$;

(iv) P^* is countably subadditive: $P^*(\cup_n A_n) \leq \sum_n P^*(A_n)$.

The others being obvious, only (iv) needs proof. For a given ϵ, choose $\mathcal{T}_0$-sets B_{nk} such that $A_n \subset \cup_k B_{nk}$ and $\sum_k P(B_{nk}) < P^*(A_n) + \epsilon 2^{-n}$, which is possible by the definition (3.1). Now $\cup_n A_n \subset \cup_{n,k} B_{nk}$, so that $P^*(\cup_n A_n) \leq \sum_{n,k} P(B_{nk}) < \sum_n P^*(A_n) + \epsilon$, and (iv) follows.[§] Of course, (iv) implies finite subadditivity.

By definition, A lies in the class $\mathcal{M}$ of P^*-measurable sets if it splits each E in 2^Ω in such a way that P^* adds for the pieces—that is, if (3.4) holds. Because of finite subadditivity, this is equivalent to

$$P^*(A \cap E) + P^*(A^c \cap E) \leq P^*(E). \tag{3.5}$$

[†]An idea which seems reasonable at first is to define $P_*(A)$ as the supremum of the sums $\sum_n P(A_n)$ for disjoint sequences of $\mathcal{T}_0$-sets in A. This will not do. For example, in the case where Ω is the unit interval, $\mathcal{T}_0$ is $\mathcal{B}_0$ (Example 2.2), and P is λ as defined by (2.12), the set N of normal numbers would have inner measure 0 because it contains no nonempty elements of $\mathcal{B}_0$; in a satisfactory theory, N will have both inner and outer measure 1.

[‡]It also turns out after the fact, that (3.3) implies that (3.4) holds for all E anyway; see Problem 3.2.

[§]Compare the proof on p. 10 that a countable union of negligible sets is negligible.

Lemma 1. *The class $\mathscr{M}$ is a field.*

Proof. It is clear that $\Omega \in \mathscr{M}$ and that $\mathscr{M}$ is closed under complementation. Suppose that $A, B \in \mathscr{M}$ and $E \subset \Omega$. Then

$$\begin{aligned} P^*(E) &= P^*(B \cap E) + P^*(B^c \cap E) \\ &= P^*(A \cap B \cap E) + P^*(A^c \cap B \cap E) \\ &\quad + P^*(A \cap B^c \cap E) + P^*(A^c \cap B^c \cap E) \\ &\geq P^*(A \cap B \cap E) \\ &\quad + P^*((A^c \cap B \cap E) \cup (A \cap B^c \cap E) \cup (A^c \cap B^c \cap E)) \\ &= P^*((A \cap B) \cap E) + P^*((A \cap B)^c \cap E), \end{aligned}$$

the inequality following by subadditivity. Hence[†] $A \cap B \in \mathscr{M}$, and $\mathscr{M}$ is a field. ■

Lemma 2. *If $A_1, A_2, \ldots$ is a finite or infinite sequence of disjoint $\mathscr{M}$-sets, then for each $E \subset \Omega$,*

$$P^*\left(E \cap \left(\bigcup_k A_k\right)\right) = \sum_k P^*(E \cap A_k). \tag{3.6}$$

Proof. Consider first the case of finitely many A_k, say n of them. For $n = 1$, there is nothing to prove. In the case $n = 2$, if $A_1 \cup A_2 = \Omega$, then (3.6) is just (3.4) with A_1 (or A_2) in the role of A. If $A_1 \cup A_2$ is smaller than Ω, split $E \cap (A_1 \cup A_2)$ by A_1 and A_1^c (or by A_2 and A_2^c) and use (3.4) and disjointness.

Assume (3.6) holds for the case of $n-1$ sets. By the case $n = 2$, together with the induction hypothesis, $P^*(E \cap (\cup_{k=1}^n A_k)) = P^*(E \cap (\cup_{k=1}^{n-1} A_k)) + P^*(E \cap A_n) = \sum_{k=1}^n P^*(E \cap A_k)$.

Thus (3.6) holds in the finite case. For the infinite case use monotonicity: $P^*(E \cap (\cup_{k=1}^\infty A_k)) \geq P^*(E \cap (\cup_{k=1}^n A_k)) = \sum_{k=1}^n P^*(E \cap A_k)$. Let $n \to \infty$, and conclude that the left side of (3.6) is greater than or equal to the right. The reverse inequality follows by countable subadditivity. ■

Lemma 3. *The class $\mathscr{M}$ is a σ-field, and P^* restricted to $\mathscr{M}$ is countably additive*

Proof. Suppose that $A_1, A_2, \ldots$ are disjoint $\mathscr{M}$-sets with union A. Since $F_n = \cup_{k=1}^n A_k$ lies in the field $\mathscr{M}$, $P^*(E) = P^*(E \cap F_n) + P^*(E \cap F_n^c)$. To the first term on the right apply (3.6), and to the second term apply monotonicity ($F_n^c \supset A^c$): $P^*(E) \geq \sum_{k=1}^n P^*(E \cap A_k) + P^*(E \cap A^c)$. Let $n \to \infty$ and use

[†]This proof does not work if (3.4) is weakened to (3.3).

(3.6) again: $P^*(E) \geq \sum_{k=1}^{\infty} P^*(E \cap A_k) + P^*(E \cap A^c) = P^*(E \cap A) + P^*(E \cap A^c)$. Hence A satisfies (3.5) and so lies in $\mathscr{M}$, which is therefore closed under the formation of countable disjoint unions.

From the fact that $\mathscr{M}$ is a field closed under the formation of countable disjoint unions it follows that $\mathscr{M}$ is a σ-field (for sets B_k in $\mathscr{M}$, let $A_1 = B_1$ and $A_k = B_k \cap B_1^c \cap \cdots \cap B_{k-1}^c$; then the A_k are disjoint $\mathscr{M}$-sets and $\bigcup_k B_k = \bigcup_k A_k \in \mathscr{M}$). The countable additivity of P^* on $\mathscr{M}$ follows from (3.6): take $E = \Omega$. ■

Lemmas 1, 2, and 3 use only the properties (i) through (iv) of P^* derived above. The next two use the specific assumption that P^* is defined via (3.1) from a probability measure P on the field $\mathscr{F}_0$.

Lemma 4. *If P^* is defined by* (3.1)*, then $\mathscr{F}_0 \subset \mathscr{M}$.*

Proof. Suppose that $A \in \mathscr{F}_0$. Given E and ϵ, choose $\mathscr{F}_0$-sets A_n such that $E \subset \bigcup_n A_n$ and $\sum_n P(A_n) \leq P^*(E) + \epsilon$. The sets $B_n = A_n \cap A$ and $C_n = A_n \cap A^c$ lie in $\mathscr{F}_0$ because it is a field. Also, $E \cap A \subset \bigcup_n B_n$ and $E \cap A^c \subset \bigcup_n C_n$; by the definition of P^* and the finite additivity of P, $P^*(E \cap A) + P^*(E \cap A^c) \leq \sum_n P(B_n) + \sum_n P(C_n) = \sum_n P(A_n) \leq P^*(E) + \epsilon$. Hence $A \in \mathscr{F}_0$ implies (3.5), and so $\mathscr{F}_0 \subset \mathscr{M}$. ■

Lemma 5. *If P^* is defined by (3.1), then*

$$P^*(A) = P(A) \quad \textit{for } A \in \mathscr{F}_0. \tag{3.7}$$

Proof. It is obvious from the definition (3.1) that $P^*(A) \leq P(A)$ for A in $\mathscr{F}_0$. If $A \subset \bigcup_n A_n$, where A and the A_n are in $\mathscr{F}_0$, then by the countable subadditivity and monotonicity of P on $\mathscr{F}_0$. $P(A) \leq \sum_n P(A \cap A_n) \leq \sum_n P(A_n)$. Hence (3.7). ■

Proof of Extension in Theorem 3.1. Suppose that P^* is defined via (3.1) from a (countably additive) probability measure P on the field $\mathscr{F}_0$. Let $\mathscr{F} = \sigma(\mathscr{F}_0)$. By Lemmas 3 and 4,[†]

$$\mathscr{F}_0 \subset \mathscr{F} \subset \mathscr{M} \subset 2^{\Omega}.$$

By (3.7), $P^*(\Omega) = P(\Omega) = 1$. By Lemma 3, P^* (which is defined on all of 2^{Ω}) restricted to $\mathscr{M}$ is therefore a probability measure there. And then P^* further restricted to $\mathscr{F}$ is clearly a probability measure on that class as well. This measure on $\mathscr{F}$ is the required extension, because by (3.7) it agrees with P on $\mathscr{F}_0$. ■

[†]In the case of Lebesgue measure, the relation is $\mathscr{B}_0 \subset \mathscr{B} \subset \mathscr{M} \subset 2^{(0,1)}$, and each of the three inclusions is strict; see Example 2.2 and Problems 3.14 and 3.21.

Uniqueness and the $\pi-\lambda$ Theorem

To prove the extension in Theorem 3.1 is unique requires some auxiliary concepts. A class $\mathscr{P}$ of subsets of Ω is a *π-system* if it is closed under the formation of finite intersections:

(π) $A, B \in \mathscr{P}$ implies $A \cap B \in \mathscr{P}$.

A class $\mathscr{L}$ is a *λ-system* if it contains Ω and is closed under the formation of complements and of finite and countable *disjoint* unions:

(λ_1) $\Omega \in \mathscr{L}$;
(λ_2) $A \in \mathscr{L}$ implies $A^c \in \mathscr{L}$;
(λ_3) $A_1, A_2, \ldots, \in \mathscr{L}$ and $A_n \cap A_m = \emptyset$ for $m \neq n$ imply $\bigcup_n A_n \in \mathscr{L}$.

Because of the disjointness condition in (λ_3), the definition of λ-system is weaker (more inclusive) than that of σ-field. In the presence of (λ_1) and (λ_2), which imply $\emptyset \in \mathscr{L}$, the countably infinite case of (λ_3) implies the finite one.

In the presence of (λ_1) and (λ_3), (λ_2) is equivalent to the condition that $\mathscr{L}$ is closed under the formation of proper differences:

(λ_2') $A, B \in \mathscr{L}$ and $A \subset B$ imply $B - A \in \mathscr{L}$.

Suppose, in fact, that $\mathscr{L}$ satisfies (λ_2) and (λ_3). If $A, B \in \mathscr{L}$ and $A \subset B$, then $\mathscr{L}$ contains B^c, the disjoint union $A \cup B^c$, and its complement $(A \cup B^c)^c = B - A$. Hence (λ_2'). On the other hand, if $\mathscr{L}$ satisfies (λ_1) and (λ_2'), then $A \in \mathscr{L}$ implies $A^c = \Omega - A \in \mathscr{L}$. Hence (λ_2).

Although a σ-field is a λ-system, the reverse is not true (in a four-point space take $\mathscr{L}$ to consist of $\emptyset, \Omega$, and the six two-point sets). But the connection is close:

Lemma 6. *A class that is both a π-system and a λ-system is a σ-field.*

Proof. The class contains Ω by (λ_1) and is closed under the formation of complements and finite intersections by (λ_2) and (π). It is therefore a field. It is a σ-field because if it contains sets A_n, then it also contains the disjoint sets $B_n = A_n \cap A_1^c \cap \cdots \cap A_{n-1}^c$ and by (λ_3) contains $\bigcup_n A_n = \bigcup_n B_n$. ■

Many uniqueness arguments depend on *Dynkin's $\pi-\lambda$ theorem*:

THEOREM 3.2
If $\mathscr{P}$ is a π-system and $\mathscr{L}$ is a λ-system, then $\mathscr{P} \subset \mathscr{L}$ implies $\sigma(\mathscr{P}) \subset \mathscr{L}$.

Proof. Let $\mathscr{L}_0$ be the λ-system generated by $\mathscr{P}$—that is, the intersection of all λ-systems containing $\mathscr{P}$. It is a λ-system, it contains $\mathscr{P}$, and it is contained in every λ-system that contains $\mathscr{P}$ (see the construction of generated σ-fields, p. 21). Thus $\mathscr{P} \subset \mathscr{L}_0 \subset \mathscr{L}$. If it can be shown that $\mathscr{L}_0$ is also a π-system, then it will follow by Lemma 6 that it is a σ-field. From the minimality of $\sigma(\mathscr{P})$ it will then follow that $\sigma(\mathscr{P}) \subset \mathscr{L}_0$, so that $\mathscr{P} \subset \sigma(\mathscr{P}) \subset \mathscr{L}_0 \subset \mathscr{L}$. Therefore, it suffices to show that $\mathscr{L}_0$ is a π-system.

For each A, let $\mathscr{L}_A$ be the class of sets B such that $A \cap B \in \mathscr{L}_0$. If A is assumed to lie in $\mathscr{P}$, or even if A is merely assumed to lie in $\mathscr{L}_0$, then $\mathscr{L}_A$ *is a* λ*-system*: Since $A \cap \Omega = A \in \mathscr{L}_0$ by the assumption, $\mathscr{L}_A$ satisfies (λ_1). If $B_1, B_2 \in \mathscr{L}_A$ and $B_1 \subset B_2$, then the λ-system $\mathscr{L}_0$ contains $A \cap B_1$ and $A \cap B_2$ and hence contains the proper difference $(A \cap B_2) - (A \cap B_1) = A \cap (B_2 - B_1)$, so that $\mathscr{L}_A$ contains $B_2 - B_1$: $\mathscr{L}_A$ satisfies (λ_2'). If B_n are disjoint $\mathscr{L}_A$-sets, then $\mathscr{L}_0$ contains the disjoint sets $A \cap B_n$ and hence contains their union $A \cap (\bigcup_n B_n)$: $\mathscr{L}_A$ satisfies (λ_3).

If $A \in \mathscr{P}$ and $B \in \mathscr{P}$, then ($\mathscr{P}$ is a π-system) $A \cap B \in \mathscr{P} \subset \mathscr{L}_0$, or $B \in \mathscr{L}_A$. Thus $A \in \mathscr{P}$ implies $\mathscr{P} \subset \mathscr{L}_A$, and since $\mathscr{L}_A$ is a λ-system, minimality gives $\mathscr{L}_0 \subset \mathscr{L}_A$.

Thus $A \in \mathscr{P}$ implies $\mathscr{L}_0 \subset \mathscr{L}_A$, or, to put it another way, $A \in \mathscr{P}$ and $B \in \mathscr{L}_0$ together imply that $B \in \mathscr{L}_A$ and hence $A \in \mathscr{L}_B$. (The key to the proof is that $B \in \mathscr{L}_A$ if and only if $A \in \mathscr{L}_B$.) This last implication means that $B \in \mathscr{L}_0$ implies $\mathscr{P} \subset \mathscr{L}_B$. Since $\mathscr{L}_B$ is a λ-system, it follows by minimality once again that $B \in \mathscr{L}_0$ implies $\mathscr{L}_0 \subset \mathscr{L}_B$. Finally, $B \in \mathscr{L}_0$ and $C \in \mathscr{L}_0$ together imply $C \in \mathscr{L}_B$, or $B \cap C \in \mathscr{L}_0$. Therefore, $\mathscr{L}_0$ is indeed a π-system. ■

Since a field is certainly a π-system, the uniqueness asserted in Theorem 3.1 is a consequence of this result:

THEOREM 3.3

Suppose that P_1 and P_2 are probability measures on $\sigma(\mathscr{P})$, where $\mathscr{P}$ is a π-system. If P_1 and P_2 agree on $\mathscr{P}$, then they agree on $\sigma(\mathscr{P})$.

Proof. Let $\mathscr{L}$ be the class of sets A in $\sigma(\mathscr{P})$ such that $P_1(A) = P_2(A)$. Clearly $\Omega \in \mathscr{L}$. If $A \in \mathscr{L}$, then $P_1(A^c) = 1 - P_1(A) = 1 - P_2(A) = P_2(A^c)$, and hence $A^c \in \mathscr{L}$. If A_n are disjoint sets in $\mathscr{L}$, then $P_1(\bigcup_n A_n) = \Sigma_n P_1(A_n) = \Sigma_n P_2(A_n) = P_2(\bigcup_n A_n)$, and hence $\bigcup_n A_n \in \mathscr{L}$. Therefore $\mathscr{L}$ is a λ-system. Since by hypothesis $\mathscr{P} \subset \mathscr{L}$ and $\mathscr{P}$ is a π-system, the $\pi - \lambda$ theorem gives $\sigma(\mathscr{P}) \subset \mathscr{L}$, as required. ■

Note that the π-λ theorem and the concept of λ-system are exactly what are needed to make this proof work: The essential property of probability measures is countable additivity, and this is a condition on countable *disjoint* unions, the only kind involved in the requirement (λ_3) in the definition of λ-system.

In this, as in many applications of the π-λ theorem, $\mathscr{L} \subset \sigma(\mathscr{P})$ and therefore $\sigma(\mathscr{P}) = \mathscr{L}$, even though the relation $\sigma(\mathscr{P}) \subset \mathscr{L}$ itself suffices for the conclusion of the theorem.

Monotone Classes

A class $\mathscr{M}$ of subsets of Ω is *monotone* if it is closed under the formation of monotone unions and intersections:

(i) $A_1, A_2, \ldots \in \mathscr{M}$ and $A_n \uparrow A$ imply $A \in \mathscr{M}$;

(ii) $A_1, A_2, \ldots \in \mathscr{M}$ and $A_n \downarrow A$ imply $A \in \mathscr{M}$.

Halmos's monotone class theorem is a close relative of the π-λ theorem but will be less frequently used in this book.

THEOREM 3.4
If $\mathscr{F}_0$ is a field and $\mathscr{M}$ is a monotone class, then $\mathscr{F}_0 \subset \mathscr{M}$ implies $\sigma(\mathscr{F}_0) \subset \mathscr{M}$.

Proof. Let $m(\mathscr{F}_0)$ be the minimal monotone class over $\mathscr{F}_0$—the intersection of all monotone classes containing $\mathscr{F}_0$. It is enough to prove $\sigma(\mathscr{F}_0) \subset m(\mathscr{F}_0)$; this will follow if $m(\mathscr{F}_0)$ is shown to be a field, because a monotone field is a σ-field.

Consider the class $\mathscr{G} = [A: A^c \in m(\mathscr{F}_0)]$. Since $m(\mathscr{F}_0)$ is monotone, so is $\mathscr{G}$. Since $\mathscr{F}_0$ is a field, $\mathscr{F}_0 \subset \mathscr{G}$, and so $m(\mathscr{F}_0) \subset \mathscr{G}$. Hence $m(\mathscr{F}_0)$ is closed under complementation.

Define $\mathscr{G}_1$ as the class of A such that $A \cup B \in m(\mathscr{F}_0)$ for all $B \in \mathscr{F}_0$. Then $\mathscr{G}_1$ is a monotone class and $\mathscr{F}_0 \subset \mathscr{G}_1$; from the minimality of $m(\mathscr{F}_0)$ follows $m(\mathscr{F}_0) \subset \mathscr{G}_1$. Define $\mathscr{G}_2$ as the class of B such that $A \cup B \in m(\mathscr{F}_0)$ for all $A \in m(\mathscr{F}_0)$. Then $\mathscr{G}_2$ is a monotone class. Now from $m(\mathscr{F}_0) \subset \mathscr{G}_1$ it follows that $A \in m(\mathscr{F}_0)$ and $B \in \mathscr{F}_0$ together imply that $A \cup B \in m(\mathscr{F}_0)$; in other words, $B \in \mathscr{F}_0$ implies that $B \in \mathscr{G}_2$. Thus $\mathscr{F}_0 \subset \mathscr{G}_2$; by minimality, $m(\mathscr{F}_0) \subset \mathscr{G}_2$, and hence $A, B \in m(\mathscr{F}_0)$ implies that $A \cup B \in m(\mathscr{F}_0)$. ■

Lebesgue Measure on the Unit Interval

Consider once again the unit interval (0, 1] together with the field $\mathscr{B}_0$ of finite disjoint unions of subintervals (Example 2.2) and the σ-field $\mathscr{B} = \sigma(\mathscr{B}_0)$ of Borel sets in (0, 1]. According to Theorem 2.2, (2.12) defines a probability measure λ on $\mathscr{B}_0$. By Theorem 3.1, λ extends to $\mathscr{B}$, the extended λ being Lebesgue measure. The probability space $((0, 1], \mathscr{B}, \lambda)$ will be the basis for much of the probability theory in the remaining sections of this chapter. A few geometric properties of λ will be considered here. Since the intervals in (0, 1] from a π-system generating $\mathscr{B}, \lambda$ is the only probability measure on $\mathscr{B}$ that assigns to each interval its length as its measure.

Some Borel sets are difficult to visualize:

EXAMPLE 3.1

Let $\{r_1, r_2, \ldots\}$ be an enumeration of the rationals in (0, 1). Suppose that ϵ is small, and choose an open interval $I_n = (a_n, b_n)$ such that $r_n \in I_n \subset (0, 1)$ and $\lambda(I_n) = b_n - a_n < \epsilon 2^{-n}$. Put $A = \bigcup_{n=1}^{\infty} I_n$. By subadditivity, $0 < \lambda(A) < \epsilon$.

Since A contains all the rationals in (0, 1), it is dense there. Thus A is an open, dense set with measure near 0. If I is an open subinterval of (0, 1), then I must intersect one of the I_n, and therefore $\lambda(A \cap I) > 0$.

If $B = (0, 1) - A$ then $1 - \epsilon < \lambda(B) < 1$. The set B contains no interval and is in fact nowhere dense [A15]. Despite this, B has measure nearly 1.

EXAMPLE 3.2

There is a set defined in probability terms that has geometric properties similar to those in the preceding example. As in Section 1, let $d_n(\omega)$ be the nth digit in the dyadic expansion of ω; see (1.7). Let $A_n = [\omega \in (0, 1]: d_i(\omega) = d_{n+i}(\omega) = d_{2n+i}(\omega), i = 1, \ldots, n]$, and let $A = \bigcup_{n=1}^{\infty} A_n$. Probabilistically, A corresponds to the event that in an infinite sequence of tosses of a coin, some finite initial segment is immediately duplicated twice over. From $\lambda(A_n) = 2^n \cdot 2^{-3n}$ it follows that $0 < \lambda(A) \leq \Sigma_{n=1}^{\infty} 2^{-2n} = \frac{1}{3}$. Again A is dense in the unit interval; its measure, less than $\frac{1}{3}$, could be made less than ϵ by requiring that some initial segment be immediately duplicated k times over with k large.

The outer measure (3.1) corresponding to λ on $\mathscr{B}_0$ is the infimum of the sums $\Sigma_n \lambda(A_n)$ for which $A_n \in \mathscr{B}_0$ and $A \subset \bigcup_n A_n$. Since each A_n is a finite disjoint union of intervals, this outer measure is

$$\lambda^*(A) = \inf \sum_n |I_n|, \tag{3.8}$$

where the infimum extends over coverings of A by intervals I_n. The notion of negligibility in Section 1 can therefore be reformulated: A is negligible if and only if $\lambda^*(A) = 0$. For A in $\mathscr{B}$, this is the same thing as $\lambda(A) = 0$. This covers the set N of normal numbers: Since the complement N^c is negligible and lies in $\mathscr{B}, \lambda(N^c) = 0$. Therefore, the Borel set N itself has probability 1: $\lambda(N) = 1$.

Completeness

This is the natural place to consider completeness, although it enters into probability theory in an essential way only in connection with the study of stochastic processes in continuous time; see Sections 37 and 38.

A probability measure space $(\Omega, \mathscr{F}, P)$ is *complete* if $A \subset B, B \in \mathscr{F}$, and $P(B) = 0$ together imply that $A \in \mathscr{F}$ (and hence that $P(A) = 0$). If $(\Omega, \mathscr{F}, P)$ is complete, then the conditions $A \in \mathscr{F}, A \triangle A' \subset B \in \mathscr{F}$, and $P(B) = 0$ together imply that $A' \in \mathscr{F}$ and $P(A') = P(A)$.

Suppose that $(\Omega, \mathscr{F}, P)$ is an arbitrary probability space. Define P^* by (3.1) for $\mathscr{F}_0 = \mathscr{F} = \sigma(\mathscr{F}_0)$, and consider the σ-field $\mathscr{M}$ of P^*-measurable sets. The arguments leading to Theorem 3.1 show that P^* restricted to $\mathscr{M}$ is a probability measure. If $P^*(B) = 0$ and $A \subset B$, then $P^*(A \cap E) + P^*(A^c \cap E) \leq P^*(B) + P^*(E) = P^*(E)$ by monotonicity, so that A satisfies (3.5) and hence lies in $\mathscr{M}$. Thus $(\Omega, \mathscr{M}, P^*)$ is a complete probability measure space. *In any probability space it is therefore possible to enlarge the σ-field and extend the measure in such a way as to get a complete space*.

Suppose that $((0, 1], \mathscr{B}, \lambda)$ is completed in this way. The sets in the completed σ-field $\mathscr{M}$ are called *Lebesgue* sets, and λ extended to $\mathscr{M}$ is still called Lebesgue measure.

Nonmeasurable Sets

There exist in (0, 1] *sets that lie outside* $\mathscr{B}$. For the construction (due to Vitali) it is convenient to use addition modulo 1 in (0, 1]. For $x, y \in (0, 1]$ take $x \oplus y$ to be $x+y$ or $x + y - 1$ according as $x+y$ lies in (0, 1] or not.[†] Put $A \oplus x = [a \oplus x\colon a \in A]$.

Let $\mathscr{L}$ be the class of Borel sets A such that $A \oplus x$ is a Borel set and $\lambda(A \oplus x) = \lambda(A)$. Then $\mathscr{L}$ is a λ-system containing the intervals, and so $\mathscr{B} \subset \mathscr{L}$ by the π-λ theorem. Thus $A \in \mathscr{B}$ implies that $A \oplus x \in \mathscr{B}$ and $\lambda(A \oplus x) = \lambda(A)$. In this sense, λ is translation-invariant.

Define x and y to be equivalent ($x \sim y$) if $x \oplus r = y$ for some rational r in (0, 1]. Let H be a subset of (0, 1] consisting of exactly one representative point from each equivalence class; such a set exists under the assumption of the axiom of choice [A8]. Consider now the countably many sets $H \oplus r$ for rational r.

These sets are disjoint, because no two distinct points of H are equivalent. (If $H \oplus r_1$ and $H \oplus r_2$ share the point $h_1 \oplus r_1 = h_2 \oplus r_2$, then $h_1 \sim h_2$; this is impossible unless $h_1 = h_2$, in which case $r_1 = r_2$.) Each point of (0, 1] lies in one of these sets, because H has a representative from each equivalence class. (If $x \sim h \in H$, then $x = h \oplus r \in H \oplus r$ for some rational r.) Thus $(0, 1] = \cup_r (H \oplus r)$, a countable disjoint union.

If H were in $\mathscr{B}$, it would follow that $\lambda(0, 1] = \Sigma_r \lambda(H \oplus r)$. This is impossible: If the value common to the $\lambda(H \oplus r)$ is 0, it leads to $1 = 0$; if the common

[†]This amounts to working in the circle group, where the translation $y \to x \oplus y$ becomes a rotation (1 is the identity). The rationals form a subgroup, and the set H defined below contains one element from each coset.

value is positive, it leads to a convergent infinite series of identical positive terms ($a + a + \cdots < \infty$ and $a > 0$). Thus H lies outside $\mathscr{B}$.

Two Impossibility Theorems†

The argument above, which uses the axiom of choice, in fact proves this: *There exists on* $2^{(0,1]}$ *no probability measure* P *such that* $P(A \oplus x) = P(A)$ *for all* $A \in 2^{(0,1]}$ *and all* $x \in (0, 1]$. In particular it is impossible to extend λ to a translation-invariant probability measure on $2^{(0,1]}$.

There is a stronger result: *There exists on* $2^{(0,1]}$ *no probability measure* P *such that* $P\{x\} = 0$ *for each* x. Since $\lambda\{x\} = 0$, this implies that it is impossible to extend λ to $2^{(0,1]}$ at all.‡

The proof of this second impossibility theorem requires the well-ordering principle (equivalent to the axiom of choice) and also the continuum hypothesis. Let S be the set of sequences $(s(1), s(2), \ldots)$ of positive integers. Then S has the power of the continuum. (Let the nth partial sum of a sequence in S be the position of the nth 1 in the nonterminating dyadic representation of a point in (0, 1]; this gives a one-to-one correspondence.) By the continuum hypothesis, the elements of S can be put in a one-to-one correspondence with the set of ordinals preceding the first uncountable ordinal. Carrying the well ordering of these ordinals over to S by means of the correspondence gives to S a well-ordering relation $\leq_w$ with the property that each element has only countably many predecessors.

For s, t in S write $s \leq t$ if $s(i) \leq t(i)$ for all $i \geq 1$. Say that t *rejects* s if $t <_w s$ and $s \leq t$; this is a transitive relation. Let T be the set of unrejected elements of S. Let V_s be the set of elements that reject s, and assume it is nonempty. If t is the first element (with respect to $\leq_w$) of V_s, then $t \in T$ (if t' rejects t, then it also rejects s, and since $t' <_w t$, there is a contradiction). Therefore, if s is rejected at all, it is rejected by an element of T.

Suppose T is countable and let $t_1, t_2, \ldots$ be an enumeration of its elements. If $t^*(k) = t_k(k) + 1$, then t^* is not rejected by any t_k and hence lies in T, which is impossible because it is distinct from each t_k. Thus T is uncountable and must by the continuum hypothesis have the power of (0, 1].

Let x be a one-to-one map of T onto (0, 1]; write the image of t as x_t. Let $A_k^i = [x_t\colon t(i) = k]$ be the image under x of the set of t in T for which $t(i) = k$. Since *t(i)* must have some value k, $\cup_{k=1}^{\infty} A_k^i = (0, 1]$. Assume that P is countably additive and choose u in S in such a way that $P(\cup_{k=1}^{u(i)} A_k^i) \geq 1 - 1/2^{i+1}$ for

†This topic may be omitted. It uses more set theory than is assumed in the rest of the book.

‡This refers to a countably additive extension, of course. If one is content with finite additivity, there *is* an extension to $2^{(0,1]}$; see Problem 3.8.

$i \geq 1$. If

$$A = \bigcap_{i=1}^{\infty} \bigcup_{k=1}^{u(i)} A_k^i = \bigcap_{i=1}^{\infty} [x_t\colon t(i) \leq u(i)] = [x_t\colon t \leq u],$$

then $P(A) > 0$. It A is shown to be countable, this will contradict the hypothesis that each singleton has probability 0.

Now, there is some t_0 in T such that $u \leq t_0$ (if $u \in T$, take $t_0 = u$; otherwise, u is rejected by some t_0 in T). If $t \leq u$ for a t in T, then $t \leq t_0$ and hence $t \leq_w t_0$ (since otherwise t_0 rejects t). This means that $[t\colon t \leq u]$ is contained in the countable set $[t\colon t \leq_w t_0]$, and A is indeed countable.

PROBLEMS

3.1. **(a)** In the proof of Theorem 3.1 the assumed finite additivity of P is used twice and the assumed countable additivity of P is used once. Where?

(b) Show by example that a finitely additive probability measure on a field may not be countably subadditive. Show in fact that if a finitely additive probability measure is countably subadditive, then it is necessarily countably additive as well.

(c) Suppose Theorem 2.1 were weakened by strengthening its hypothesis to the assumption that $\mathcal{F}$ is a σ-field. Why would this weakened result not suffice for the proof of Theorem 3.1?

3.2. Let P be a probability measure on a field $\mathcal{F}_0$ and for every subset A of Ω define $P^*(A)$ by (3.1). Denote also by P the extension (Theorem 3.1) of P to $\mathcal{F} = \sigma(\mathcal{F}_0)$.

(a) Show that

$$P^*(A) = \inf[P(B)\colon A \subset B, B \in \mathcal{F}] \tag{3.9}$$

and (see (3.2))

$$P_*(A) = \sup[P(C)\colon C \subset A, C \in \mathcal{F}], \tag{3.10}$$

and show that the infimum and supremum are always achieved.

(b) Show that A is P^*-measurable if and only if $P_*(A) = P^*(A)$.

(c) The outer and inner measures associated with a probability measure P on a σ-field $\mathcal{F}$ are usually *defined* by (3.9) and (3.10). Show that (3.9) and (3.10) are the same as (3.1) and (3.2) with $\mathcal{F}$ in the role of $\mathcal{F}_0$.

3.3. 2.13, 2.15, 3.2 ↑ For the following examples, describe P^* as defined by (3.1) and $\mathscr{M} = \mathscr{M}(P^*)$ as defined by the requirement (3.4). Sort out the cases in which P^* fails to agree with P on $\mathscr{F}_0$ and explain why.

(a) Let $\mathscr{F}_0$ consist of the sets $\emptyset, \{1\}, \{2,3\}$, and $\Omega = \{1,2,3\}$, and define probability measures P_1 and P_2 on $\mathscr{F}_0$ by $P_1\{1\} = 0$ and $P_2\{2,3\} = 0$, Note that $\mathscr{M}(P_1^*)$ and $\mathscr{M}(P_2^*)$ differ.

(b) Suppose that Ω is countably infinite, let $\mathscr{F}_0$ be the field of finite and cofinite sets, and take $P(A)$ to be 0 or 1 as A is finite or confinite.

(c) The same, but suppose that Ω is uncountable.

(d) Suppose that Ω is uncountable, let $\mathscr{F}_0$ consist of the countable and the cocountable sets, and take $P(A)$ to be 0 or 1 as A is countable or cocountable.

(e) The probability in Problem 2.15.

(f) Let $P(A) = I_A(\omega_0)$ for $A \in \mathscr{F}_0$, and assume $\{\omega_0\} \in \sigma(\mathscr{F}_0)$.

3.4. Let f be a strictly increasing, strictly concave function on $[0,\infty)$ satisfying $f(0) = 0$. For $A \subset (0,1]$, define $P^*(A) = f(\lambda^*(A))$. Show that P^* is an outer measure in the sense that it satisfies $P^*(\emptyset) = 0$ and is nonnegative, monotone, and countably subadditive. Show that A lies in $\mathscr{M}$ (defined by the requirement (3.4)) if and only if $\lambda^*(A)$ or $\lambda^*(A^c)$ is 0. Show that P^* does not arise from the definition (3.1) for any probability measure P on any field $\mathscr{F}_0$.

3.5. Let Ω be the unit square $[(x,y)\colon 0 < x,y \le 1]$, let $\mathscr{F}$ be the class of sets of the form $[(x,y)\colon x \in A, 0 < y \le 1]$, where $A \in \mathscr{B}$, and let P have value $\lambda(A)$ at this set. Show that $(\Omega, \mathscr{F}, P)$ is a probability measure space. Show for $A = [(x,y)\colon 0 < x \le 1, y = \frac{1}{2}]$ that $P_*(A) = 0$ and $P^*(A) = 1$.

3.6. Let P be a *finitely* additive probability measure on a field $\mathscr{F}_0$. For $A \subset \Omega$, in analogy with (3.1) define

$$P^\circ(A) = \inf \sum_n P(A_n), \tag{3.11}$$

where now the infimum extends over all *finite* sequences of $\mathscr{F}_0$-sets A_n satisfying $A \subset \bigcup_n A_n$. (If countable coverings are allowed, everything is different. It can happen that $P^\circ(\Omega) = 0$; see Problem 3.3(e).) Let $\mathscr{M}^\circ$ be the class of sets A such that $P^\circ(E) = P^\circ(A \cap E) + P^\circ(A^c \cap E)$ for all $E \subset \Omega$.

(a) Show that $P^\circ(\emptyset) = 0$ and that P° is nonnegative, monotone, and *finitely* subadditive. Using these four properties of P°, prove: Lemma 1°: $\mathscr{M}^\circ$ is a field. Lemma 2°: If $A_1, A_2, \ldots$ is a *finite* sequence of disjoint $\mathscr{M}^\circ$-sets, then for each $E \subset \Omega$,

$$P^\circ\left(E \cap \left(\bigcup_k A_k\right)\right) = \sum_k P^\circ(E \cap A_k). \tag{3.12}$$

Lemma 3°: P° restricted to the field $\mathscr{M}^\circ$ is *finitely* additive.

(b) Show that if P° is defined by (3.11) (finite coverings), then: Lemma 4°: $\mathscr{F}_0 \subset \mathscr{M}^\circ$. Lemma 5°: $P^\circ(A) = P(A)$ for $A \in \mathscr{F}_0$.

(c) Define $P_\circ(A) = 1 - P^\circ(A^c)$. Prove that if $E \subset A \in \mathscr{F}_0$, then

$$P_\circ(E) = P(A) - P^\circ(A - E). \tag{3.13}$$

3.7. 2.7 3.6 ↑ Suppose that H lies outside the field $\mathscr{F}_0$, and let $\mathscr{F}_1$ be the field generated by $\mathscr{F}_0 \cup \{H\}$, so that $\mathscr{F}_1$ consists of the sets $(H \cap A) \cup (H^c \cap B)$ with $A, B \in \mathscr{F}_0$. The problem is to show that a finitely additive probability measure P on $\mathscr{F}_0$ has a finitely additive extension to $\mathscr{F}_1$. Define Q on $\mathscr{F}_1$ by

$$Q((H \cap A) \cup (H^c \cap B)) = P^\circ(H \cap A) + P_\circ(H^c \cap B) \tag{3.14}$$

for $A, B \in \mathscr{F}_0$.

(a) Show that the definition is consistent.

(b) Shows that Q agrees with P on $\mathscr{F}_0$.

(c) Show that Q is finitely additive on $\mathscr{F}_1$. Show that $Q(H) = P^\circ(H)$.

(d) Define Q' by interchanging the roles of P° and $P_\circ$ on the right in (3.14). Show that Q' is another finitely additive extension of P to $\mathscr{F}_1$. The same is true of any convex combination Q'' of Q and Q'. Show that $Q''(H)$ can take any value between $P_\circ(H)$ and $P^\circ(H)$.

3.8. ↑ Use Zorn's lemma to prove a theorem of Tarski: A finitely additive probability measure on a field has a finitely additive extension to the field of all subsets of the space.

3.9. ↑

(a) Let P be a (countably additive) probability measure on a σ-field $\mathscr{F}$. Suppose that $H \notin \mathscr{F}$, and let $\mathscr{F}_1 = \sigma(\mathscr{F} \cup \{H\})$. By adapting the ideas in Problem 3.7, show that P has a countably additive extension from $\mathscr{F}$ to $\mathscr{F}_1$.

(b) It is tempting to go on and use Zorn's lemma to extend P to a completely additive probability measure on the σ-field of all subsets of Ω. Where does the obvious proof break down?

3.10. 2.17 3.2 ↑ As shown in the text, a probability measure space $(\Omega, \mathscr{F}, P)$ has a complete extension—that is, there exists a complete probability measure space $(\Omega, \mathscr{F}_1, P_1)$ such that $\mathscr{F} \subset \mathscr{F}_1$ and P_1 agrees with P on $\mathscr{F}$.

(a) Suppose that $(\Omega, \mathscr{F}_2, P_2)$ is a second complete extension. Show by an example in a space of two points that P_1 and P_2 need not agree on the σ-field $\mathscr{F}_1 \cap \mathscr{F}_2$.

(b) There is, however, a unique minimal complete extension: Let $\mathscr{F}^+$ consist of the sets A for which there exist $\mathscr{F}$-sets B and C such that $A \triangle B \subset C$ and $P(C) = 0$. Show that $\mathscr{F}^+$ is a σ-field.

For such a set A define $P^+(A) = P(B)$. Show that the definition is consistent, that P^+ is a probability measure on $\mathscr{F}^+$, and that $(\Omega, \mathscr{F}^+, P^+)$ is complete. Show that, if $(\Omega, \mathscr{F}_1, P_1)$ is any complete extension of $(\Omega, \mathscr{F}, P)$, then $\mathscr{F}^+ \subset \mathscr{F}_1$ and P_1 agrees with P^+ on $\mathscr{F}^+$; $(\Omega, \mathscr{F}^+, P^+)$ is the *completion* of $(\Omega, \mathscr{F}, P)$.

(c) Show that $A \in \mathscr{F}^+$ if and only if $P_*(A) = P^*(A)$, where P_* and P^* are defined by (3.9) and (3.10), and that $P^+(A) = P_*(A) = P^*(A)$ in this case. Thus the complete extension constructed in the text is exactly *the* completion.

3.11. **(a)** Show that a λ-system satisfies the conditions
(λ_4) $A, B \in \mathscr{L}$ and $A \cap B = \varnothing$ imply $A \cup B \in \mathscr{L}$,
(λ_5) $A_1, A_2, \ldots \in \mathscr{L}$ and $A_n \uparrow A$ imply $A \in \mathscr{L}$,
(λ_6) $A_1, A_2, \ldots \in \mathscr{L}$ and $A_n \downarrow A$ imply $A \in \mathscr{L}$.

(b) Show that $\mathscr{L}$ is a λ-system if and only if it satisfies (λ_1), (λ_2'), and (λ_5). (Sometimes these conditions, with a redundant (λ_4), are taken as the definition.

3.12. 2.5 3.11 $\uparrow$

(a) Show that if $\mathscr{P}$ is a π-system, then the minimal λ-system over $\mathscr{P}$ coincides with $\sigma(\mathscr{P})$.

(b) Let $\mathscr{P}$ be a π-system and $\mathscr{M}$ a monotone class. Show that $\mathscr{P} \subset \mathscr{M}$ does not imply $\sigma(\mathscr{P}) \subset \mathscr{M}$.

(c) Deduce the $\pi-\lambda$ theorem from the monotone class theorem by showing directly that, if a λ-system $\mathscr{L}$ contains a π-system $\mathscr{L}$, then $\mathscr{L}$ also contains the field generated by $\mathscr{P}$.

3.13. 2.5 $\uparrow$

(a) Suppose that $\mathscr{F}_0$ is a field and P_1 and P_2 are probability measures on $\sigma(\mathscr{F}_0)$. Show by the monotone class theorem that if P_1 and P_2 agree on $\mathscr{F}_0$, then they agree on $\sigma(\mathscr{F}_0)$.

(b) Let $\mathscr{F}_0$ be the smallest field over the π-system $\mathscr{P}$. Show by the inclusion–exclusion formula that probability measures agreeing on $\mathscr{P}$ must agree also on $\mathscr{F}_0$. Now deduce Theorem 3.3 from part (a).

3.14. 1.5 2.22 $\uparrow$ Prove the existence of a Lebesgue set of Lebesgue measure 0 that is not a Borel set.

3.15. 1.3, 3.6 3.14 $\uparrow$ The *outer content* of a set A in (0, 1] is $c^*(A) = \inf \sum_n |I_n|$, where the infimum extends over *finite* coverings of A by intervals I_n. Thus A is trifling in the sense of Problem 1.3 if and only if $c^*(A) = 0$. Define *inner content* by $c_*(A) = 1 - c^*(A^c)$. Show that $c_*(A) = \sup \Sigma_n |I_n|$, where the supremum extends over finite disjoint unions of intervals I_n contained in A (of course the analogue for λ_* fails). Show that $c_*(A) \leq c^*(A)$; if the two are equal, their common value is taken as the *content* $c(A)$ of A, which is then *Jordan measurable*. Connect all this with Problem 3.6.

Show that $c^*(A) = c^*(A^-)$, where A^- is the closure of A (the analogue for λ^* fails).

A trifling set is Jordan measurable. Find (Problem 3.14) a Jordan measurable set that is not a Borel set.

Show that $c_*(A) \le \lambda_*(A) \le \lambda^*(A) \le c^*(A)$. What happens in this string of inequalities if A consists of the rationals in $(0, \frac{1}{2}]$ together with the irrationals in $(\frac{1}{2}, 1]$?

3.16. 1.5 ↑ Deduce directly by countable additivity that the Cantor set has Lebesgue measure 0.

3.17. From the fact that $\lambda(x \oplus A) = \lambda(A)$, deduce that sums and differences of normal numbers may be nonnormal.

3.18. Let H be the nonmeasurable set constructed at the end of the section.

(a) Show that, if A is a Borel set and $A \subset H$, then $\lambda(A) = 0$—that is, $\lambda_*(H) = 0$.

(b) Show that, if $\lambda^*(E) > 0$, then E contains a nonmeasurable subset.

3.19. The aim of this problem is the construction of a Borel set A in (0, 1) such that $0 < \lambda(A \cap G) < \lambda(G)$ for every nonempty open set G in (0, 1).

(a) It is shown in Example 3.1 how to construct a Borel set of positive Lebesgue measure that is nowhere dense. Show that every interval contains such a set.

(b) Let $\{I_n\}$ be an enumeration of the open intervals in (0, 1) with rational endpoints. Construct disjoint, nowhere dense Borel sets $A_1, B_1, A_2, B_2, \ldots$ of positive Lebesgue measure such that $A_n \cup B_n \subset I_n$.

(c) Let $A = \bigcup_k A_k$. A nonempty open G in (0, 1) contains some I_n. Show that $0 < \lambda(A_n) \le \lambda(A \cap G) < \lambda(A \cap G) + \lambda(B_n) \le \lambda(G)$.

3.20. ↑ There is no Borel set A in (0, 1) such that $a\lambda(I) \le \lambda(A \cap I) \le b\lambda(I)$ for every open interval I in (0, 1), where $0 < a \le b < 1$. In fact prove:

(a) If $\lambda(A \cap I) \le b\lambda(I)$ for all I and if $b < 1$, then $\lambda(A) = 0$. *Hint:* Choose an open G such that $A \subset G \subset (0, 1)$ and $\lambda(G) < b^{-1}\lambda(A)$; represent G as a disjoint union of intervals and obtain a contradiction.

(b) If $a\lambda(I) \le \lambda(A \cap I)$ for all I and if $a > 0$, then $\lambda(A) = 1$.

3.21. Show that not every subset of the unit interval is a Lebesgue set. *Hint:* Show that λ^* is translation-invariant on $2^{(0,1]}$; then use the first impossibility theorem (p. 45). Or use the second impossibility theorem.

SECTION 4 DENUMERABLE PROBABILITIES

Complex probability ideas can be made clear by the systematic use of measure theory, and probabilistic ideas of extramathematical origin, such as independence, can illuminate problems of purely mathematical interest. It is to this reciprocal exchange that measure-theoretic probability owes much of its interest.

The results of this section concern infinite sequences of events in a probability space.† They will be illustrated by examples in the *unit interval*. By this will always be meant the triple $(\Omega, \mathscr{F}, P)$ for which Ω is (0, 1], $\mathscr{F}$ is the σ-field $\mathscr{B}$ of Borel sets there, and *P(A)* is for A in $\mathscr{F}$ the Lebesgue measure $\lambda(A)$ of A. This is the space appropriate to the problems of Section 1, which will be pursued further. The definitions and theorems, as opposed to the examples, apply to *all* probability spaces. The unit interval will appear again and again in this chapter, and it is essential to keep in mind that there are many other important spaces to which the general theory will be applied later.

General Formulas

The formulas (2.5) through (2.11) will be used repeatedly. The sets involved in such formulas lie in the basic σ-field $\mathscr{F}$ by hypothesis. Any probability argument starts from given sets assumed (often tacitly) to lie in $\mathscr{F}$; further sets constructed in the course of the argument must be shown to lie in $\mathscr{F}$ as well, but it is usually quite clear how to do this.

If $P(A) > 0$, the *conditional probability* of B given A is defined in the usual way as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{4.1}$$

There are the chain-rule formulas

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A), \\ P(A \cap B \cap C) &= P(A)P(B|A)P(C|A \cap B), \end{aligned} \tag{4.2}$$

If $A_1, A_2, \ldots$ partition Ω, then

$$P(B) = \sum_n P(A_n)P(B|A_n). \tag{4.3}$$

Note that for fixed A the function $P(B|A)$ defines a probability measure as B varies over $\mathscr{F}$.

If $P(A_n) \equiv 0$, then by subadditivity $P(\bigcup_n A_n) = 0$. If $P(A_n) \equiv 1$, then $\bigcap_n A_n$ has complement $\bigcup_n A_n^c$ of probability 0. This gives two facts used over and over again:

If $A_1, A_2, \ldots$ are sets of probability 0, so is $\bigcup_n A_n$. If $A_1, A_2, \ldots$ are sets of probability 1, so is $\bigcap_n A_n$.

†They come under what Borel in his first paper on the subject (see the footnote on p. 9) called *probabilités dénombrables*; hence the section heading.

Limit Sets

For a sequence $A_1, A_2, \ldots$ of sets, define a set

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \tag{4.4}$$

and a set

$$\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k. \tag{4.5}$$

These sets[†] are the *limits superior* and *inferior* of the sequence $\{A_n\}$. They lie in $\mathscr{F}$ if all the A_n do. Now ω lies in (4.4) if and only if for each n there is some $k \geq n$ for which $\omega \in A_k$; in other words, ω lies in (4.4) if and only if it lies in *infinitely many* of the A_n. In the same way, ω lies in (4.5) if and only if there is some n such that $\omega \in A_k$ for all $k \geq n$; in other words, ω lies in (4.5) if and only if it lies in *all but finitely many* of the A_n.

Note that $\bigcap_{k=n}^{\infty} A_k \uparrow \liminf_n A_n$ and $\bigcup_{k=n}^{\infty} A_k \downarrow \limsup_n A_n$. For every m and n, $\bigcap_{k=m}^{\infty} A_k \subset \bigcup_{k=n}^{\infty} A_k$, because for $i \geq \max\{m, n\}$, A_i contains the first of these sets and is contained in the second. Taking the union over m and the intersection over n shows that (4.5) is a subset of (4.4). But this follows more easily from the observation that if ω lies in all but finitely many of the A_n then of course it lies in infinitely many of them. Facts about limits inferior and superior can usually be deduced from the logic they involve more easily than by formal set-theoretic manipulations.

If (4.4) and (4.5) are equal, write

$$\lim_n A_n = \liminf_n A_n = \limsup_n A_n. \tag{4.6}$$

To say that A_n has limit A, written $A_n \to A$, means that the limits inferior and superior do coincide and in fact coincide with A. Since $\liminf_n A_n \subset \limsup_n A_n$ always holds, to check whether $A_n \to A$ is to check whether $\limsup_n A_n \subset A \subset \liminf_n A_n$. From $A_n \in \mathscr{F}$ and $A_n \to A$ follows $A \in \mathscr{F}$.

EXAMPLE 4.1

Consider the functions $d_n(\omega)$ defined on the unit interval by the dyadic expansion (1.7), and let $l_n(\omega)$ be the length of the run of 0's starting at $d_n(\omega)$: $l_n(\omega) = k$ if $d_n(\omega) = \cdots = d_{n+k-1}(\omega) = 0$ and $d_{n+k}(\omega) = 1$; here $l_n(\omega) = 0$ if $d_n(\omega) = 1$. Probabilities can be computed by (1.10). Since $[\omega\colon l_n(\omega) = k]$ is a union of

[†]See Problems 4.1 and 4.2 for the analogy between set-theoretic and numerical limits superior and inferior.

2^{n-1} disjoint intervals of length 2^{-n-k}, it lies in $\mathscr{F}$ and has probability 2^{-k-1}. Therefore, $[\omega\colon l_n(\omega) \geq r] = [\omega\colon d_i(\omega) = 0, n \leq i < n + r]$ lies also in $\mathscr{F}$ and has probability $\Sigma_{k \geq r} 2^{-k-1}$:

$$P[\omega\colon l_n(\omega) \geq r] = 2^{-r}. \tag{4.7}$$

If A_n is the event in (4.7), then (4.4) is the set of ω such that $l_n(\omega) \geq r$ for infinitely many n, or, n being regarded as a time index, such that $l_n(\omega) \geq r$ *infinitely often*.

Because of the theory of Sections 2 and 3, statements like (4.7) are valid in the sense of ordinary mathematics, and using the traditional language of probability—"heads," "runs," and so on—does not change this.

When n has the role of time, (4.4) is frequently written

$$\limsup nA_n = [A_n \text{ i.o.}], \tag{4.8}$$

where "i.o." stands for "infinitely often."

THEOREM 4.1

(i) *For each sequence* $\{A_n\}$,

$$P\left(\liminf_n A_n\right) \leq \inf_n P(A_n) \tag{4.9}$$

$$\leq \limsup_n P(A_n) \leq P\left(\limsup_n A_n\right).$$

(ii) *If* $A_n \to A$, *then* $P(A_n) \to P(A)$.

Proof. Clearly (ii) follows from (i). As for (i), if $B_n = \cap_{k=n}^{\infty} A_k$ and $C_n = \cup_{k=n}^{\infty} A_k$, then $B_n \uparrow \liminf_n A_n$ and $C_n \downarrow \limsup_n A_n$, so that, by parts (i) and (ii) of Theorem 2.1, $P(A_n) \geq P(B_n) \to P(\liminf_n A_n)$ and $P(A_n) \leq P(C_n) \to P(\limsup_n A_n)$. ■

EXAMPLE 4.2

Define $l_n(\omega)$ as in Example 4.1, and let $A_n = [\omega\colon l_n(\omega) \geq r]$ for fixed r. By (4.7) and (4.9), $P[\omega\colon l_n(\omega) \geq r \text{ i.o.}] \geq 2^{-r}$. Much stronger results will be proved later.

Independent Events

Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$. (Sometimes an unnecessary *mutually* is put in front of *independent*.) For events of positive

probability, this is the same thing as requiring $P(B|A) = P(B)$ or $P(A|B) = P(A)$. More generally, a finite collection $A_1, \ldots, A_n$ of events is independent if

$$P(A_{k_1} \cap \cdots \cap A_{k_j}) = P(A_{k_1}) \cdots P(A_{k_j}) \tag{4.10}$$

for $2 \le j \le n$ and $1 \le k_1 < \cdots < k_j \le n$. Reordering the sets clearly has no effect on the condition for independence, and a subcollection of independent events is also independent. An infinite (perhaps uncountable) collection of events is defined to be independent in each of its finite subcollections is.

If $n = 3$, (4.10) imposes for $j = 2$ the three constraints

$$P(A_1 \cap A_2) = P(A_1)P(A_2), \quad P(A_1 \cap A_3) = P(A_1)P(A_3), \quad P(A_2 \cap A_3) = P(A_2)P(A_3), \tag{4.11}$$

and for $j = 3$ the single constraint

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3). \tag{4.12}$$

EXAMPLE 4.3

Consider in the unit interval the events $B_{uv} = [\omega\colon d_u(\omega) = d_v(\omega)]$—the uth and vth tosses agree—and let $A_1 = B_{12}, A_2 = B_{13}, A_3 = B_{23}$. Then A_1, A_2, A_3 are *pairwise* independent in the sense that (4.11) holds (the two sides of each equation being $\frac{1}{4}$). But since $A_1 \cap A_2 \subset A_3$, (4.12) does *not* hold (the left side is $\frac{1}{4}$ and the right is $\frac{1}{8}$), and the events are not independent.

EXAMPLE 4.4

In the discrete space $\Omega = \{1, \ldots, 6\}$ suppose each point has probability $\frac{1}{6}$ (a fair die is rolled). If $A_1 = \{1, 2, 3, 4\}$ and $A_2 = A_3 = \{4, 5, 6\}$, then (4.12) holds but none of the equations in (4.11) do. Again the events are not independent.

Independence requires that (4.10) hold for each $j = 2, \ldots, n$ and each choice of $k_1, \ldots, k_j$, a total of $\Sigma_{j=2}^{n} \binom{n}{j} = 2^n - 1 - n$ constraints. This requirement can be stated in a different way: If $B_1, \ldots, B_n$ are sets such that for each $i = 1, \ldots, n$ either $B_i = A_i$ or $B_i = \Omega$, then

$$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2) \cdots P(B_n). \tag{4.13}$$

The point is that if $B_i = \Omega$, then B_i can be ignored in the intersection on the left and the factor $P(B_i) = 1$ can be ignored in the product on the right. For example, replacing A_2 by Ω reduces (4.12) to the middle equation in (4.11).

From the assumed independence of certain sets it is possible to deduce the independence of other sets.

EXAMPLE 4.5

On the unit interval the events $H_n = [\omega\colon d_n(\omega) = 0], n = 1, 2, \ldots$, are independent, the two sides of (4.10) having in this case value 2^{-j}. It seems intuitively clear that from this should follow the independence, for example, of $[\omega\colon d_2(\omega) = 0] = H_2$ and $[\omega\colon d_1(\omega) = 0, d_3(\omega) = 1] = H_1 \cap H_3^c$, since the two events involve disjoint sets of times. Further, any sets A and B depending, respectively, say, only on even and on odd times (like $[\omega\colon d_{2n}(\omega) = 0$ i.o.] and $[\omega\colon d_{2n+1}(\omega) = 1$ i.o.]) ought also to be independent. This raises the general question of what it means for A to depend only on even times. Intuitively, it requires that knowing which ones among $H_2, H_4, \ldots$ occurred entails knowing whether or not A occurred—that is, it requires that the sets $H_2, H_4, \ldots$ "determine" A. The set-theoretic form of this requirement is that A is to lie in the σ-field generated by $H_2, H_4, \ldots$. From $A \in \sigma(H_2, H_4, \ldots)$ and $B \in \sigma(H_1, H_3, \ldots)$ it ought to be possible to deduce the independence of A and B.

The next theorem and its corollaries make such deductions possible. Define *classes* $\mathscr{A}_1, \ldots, \mathscr{A}_n$ in the basic σ-field $\mathscr{F}$ to be independent if for each choice of A_i from $\mathscr{A}_i, i = 1, \ldots, n$, the events $A_1, \ldots, A_n$ are independent. This is the same as requiring that (4.13) hold whenever for each $i, 1 \le i \le n$, either $B_i \in \mathscr{A}_i$ or $B_i = \Omega$.

THEOREM 4.2

If $\mathscr{A}_1, \ldots, \mathscr{A}_n$ are independent and each $\mathscr{A}_i$ is a π-system, then $\sigma(\mathscr{A}_1), \ldots, \sigma(\mathscr{A}_n)$ are independent.

Proof. Let $\mathscr{B}_i$ be the class $\mathscr{A}_i$ augmented by Ω (which may be an element of $\mathscr{A}_i$ to start with). Then each $\mathscr{B}_i$ is a π-system, and by the hypothesis of independence, (4.13) holds if $B_i \in \mathscr{B}_i, i = 1, \ldots, n$. For fixed sets $B_2, \ldots, B_n$ lying respectively in $\mathscr{B}_2, \ldots, \mathscr{B}_n$, let $\mathscr{L}$ be the class of $\mathscr{F}$-sets B_1 for which (4.13) holds. Then $\mathscr{L}$ is a λ-system containing the π-system $\mathscr{B}_1$ and hence (Theorem 3.2) containing $\sigma(\mathscr{B}_1) = \sigma(\mathscr{A}_1)$. Therefore, (4.13) holds if $B_1, B_2, \ldots, B_n$ lie respectively in $\sigma(\mathscr{A}_1), \mathscr{B}_2, \ldots, \mathscr{B}_n$, which means that $\sigma(\mathscr{A}_1), \mathscr{A}_2, \ldots, \mathscr{A}_n$ are independent. Clearly the argument goes through if 1 is replaced by any of the indices $2, \ldots, n$.

From the independence of $\sigma(\mathscr{A}_1), \mathscr{A}_2, \ldots, \mathscr{A}_n$ now follows that of $\sigma(\mathscr{A}_1), \sigma(\mathscr{A}_2), \mathscr{A}_3, \ldots, \mathscr{A}_n$, and so on. ■

If $\mathscr{A} = \{A_1, \ldots, A_k\}$ is finite, then each A in $\sigma(\mathscr{A})$ can be expressed by a "formula" such as $A = A_2 \cap A_5^c$ or $A = (A_2 \cap A_7) \cup (A_3 \cap A_7^c \cap A_8)$. If $\mathscr{A}$ is infinite, the sets in $\sigma(\mathscr{A})$ may be very complicated; the way to make precise the idea that the elements of $\mathscr{A}$ "determine" A is not to require formulas, but simply to require that A lie in $\sigma(\mathscr{A})$.

Independence for an infinite collection of classes is defined just as in the finite case: $[\mathscr{A}_\theta\colon \theta \in \Theta]$ is independent if the collection $[A_\theta\colon \theta \in \Theta]$ of sets is independent for each choice of A_θ from $\mathscr{A}_\theta$. This is equivalent to the independence of each finite subcollection $\mathscr{A}_{\theta_1}, \ldots, \mathscr{A}_{\theta_n}$ of classes, because of the way independence for infinite classes of sets is defined in terms of independence for finite classes. Hence Theorem 4.2 has an immediate consequence:

Corollary 1. *If $A_\theta, \theta \in \Theta$, are independent and each $\mathscr{A}_\theta$ is a π-system, then $\sigma(\mathscr{A}_\theta), \theta \in \Theta$, are independent.*

Corollary 2. *Suppose that the array*

$$\begin{array}{llll} A_{11} & A_{12} & \ldots \\ A_{21} & A_{22} & \ldots \\ \vdots & \vdots \end{array} \tag{4.14}$$

of events is independent; here each row is a finite or infinite sequence, and there are finitely or infinitely many rows. If $\mathscr{F}_i$ is the σ-field generated by the i th row, then $\mathscr{F}_1, \mathscr{F}_2, \ldots$ are independent.

Proof. If $\mathscr{A}_i$ is the class of all finite intersections of elements of the ith row of (4.14), then $\mathscr{A}_i$ is a π-system and $\sigma(\mathscr{A}_i) = \mathscr{F}_i$. Let I be a finite collection of indices (integers), and for each i in I let J_i be a finite collection of indices. Consider for $i \in I$ the element $C_i = \cap_{j \in J_i} A_{ij}$ of $\mathscr{A}_i$. Since every finite subcollection of the array (4.14) is independent (the intersections and products here extend over $i \in I$ and $j \in J_i$),

$$\begin{aligned} P\left(\bigcap_i C_i\right) = P\left(\bigcap_i \bigcap_j A_{ij}\right) &= \prod_i \prod_j P(A_{ij}) = \prod_i P(\cap_j A_{ij}) \\ &= \prod_i P(C_i). \end{aligned}$$

It follows that the classes $\mathscr{A}_1, \mathscr{A}_2, \ldots$ are independent, so that Corollary 1 applies. ■

Corollary 2 implies the independence of the events discussed in Example 4.5. The array (4.14) in this case has two rows:

$$\begin{array}{llll} H_2 & H_4 & H_6 & \cdots \\ H_1 & H_3 & H_5 & \cdots \end{array}$$

Theorem 4.2 also implies, for example, that for independent $A_1, \ldots, A_n$,

$$P(A_1^c \cap \cdots \cap A_k^c \cap A_{k+1} \cap \cdots \cap A_n) = P(A_1^c) \cdots P(A_k^c) P(A_{k+1}) \cdots P(A_n). \tag{4.15}$$

To prove this, let $\mathscr{A}_i$ consist of A_i alone; of course, $A_i^c \in \sigma(\mathscr{A}_i)$. In (4.15) any subcollection of the A_i could be replaced by their complements.

EXAMPLE 4.6

Consider as in Example 4.3 the events B_{uv} that, in a sequence of tosses of a fair coin, the uth and vth outcomes agree. Let $\mathscr{A}_1$ consist of the events B_{12} and B_{13}, and let $\mathscr{A}_2$ consist of the event B_{23}. Since these three events are pairwise independent, the classes $\mathscr{A}_1$ and $\mathscr{A}_2$ are independent. Since $B_{23} = (B_{12} \triangle B_{13})^c$ lies in $\sigma(\mathscr{A}_1)$, however, $\sigma(\mathscr{A}_1)$ and $\sigma(\mathscr{A}_2)$ are not independent. The trouble is that $\mathscr{A}_1$ is not a π-system, which shows that this condition in Theorem 4.2 is essential.

EXAMPLE 4.7

If $\mathscr{A} = \{A_1, A_2, \ldots\}$ is a finite or countable partition of Ω and $P(B|A_i) = p$ for each A_i of positive probability, then $P(B) = p$ and B is independent of $\sigma(\mathscr{A})$: If Σ' denotes summation over those i for which $P(A_i) > 0$, then $P(B) = \Sigma' P(A_i \cap B) = \Sigma' P(A_i) p = p$, and so B is independent of each set in the π-system $\mathscr{A} \cup \{\emptyset\}$.

Subfields

Theorem 4.2 involves a number of σ-fields at once, which is characteristic of probability theory; measure theory not directed toward probability usually involves only one all-embracing σ-field $\mathscr{F}$. In proability, σ-fields in $\mathscr{F}$—that is, sub-σ-fields of $\mathscr{F}$—play an important role. To understand their function it helps to have an informal, intuitive way of looking at them.

A subclass $\mathscr{A}$ of $\mathscr{F}$ corresponds heuristically to *partial information*. Imagine that a point ω is drawn from Ω according to the probabilities given by P: ω lies in A with probability $P(A)$. Imagine also an observer who does not know which ω it is that has been drawn but who does know for each A in $\mathscr{A}$ whether $\omega \in A$ or $\omega \notin A$—that is, who does not know ω but does know the value of $I_A(\omega)$ for each A in $\mathscr{A}$. Identifying this partial information with the class $\mathscr{A}$ itself will illuminate the connection between various measure-theoretic concepts and the premathematical ideas lying behind them.

The set B is by definition independent of the class $\mathscr{A}$ if $P(B|A) = P(B)$ for all sets A in $\mathscr{A}$ for which $P(A) > 0$. Thus if B is independent of $\mathscr{A}$, then the

observer's probability for B is $P(B)$ even after he has received the information in $\mathscr{A}$; in this case $\mathscr{A}$ contains no information about B. The point of Theorem 4.2 is that this remains true even if the observer is given the information in $\sigma(\mathscr{A})$, provided that $\mathscr{A}$ is a π-system. It is to be stressed that here *information*, like *observer* and *know*, is an informal, extramathematical term (in particular, it is not information in the technical sense of entropy).

The notion of partial information can be looked at in terms of partitions. Say that points ω and ω' are *$\mathscr{A}$-equivalent* if, for every A in $\mathscr{A}$, ω and ω' lie either both in A or both in A^c—that is, if

$$I_A(\omega) = I_A(\omega'), \qquad A \in \mathscr{A}. \tag{4.16}$$

This relation partitions Ω into sets of equivalent points; call this the *$\mathscr{A}$ partition*.

EXAMPLE 4.8

If ω and ω' are $\sigma(\mathscr{A})$-equivalent, then certainly they are $\mathscr{A}$-equivalent. For fixed ω and ω', the class of A such that $I_A(\omega) = I_A(\omega')$ is a σ-field; if ω and ω' are $\mathscr{A}$-equivalent, then this σ-field contains $\mathscr{A}$ and hence $\sigma(\mathscr{A})$, so that ω and ω' are also $\sigma(\mathscr{A})$-equivalent. Thus $\mathscr{A}$-equivalence and $\sigma(\mathscr{A})$-equivalence are the same thing, and the $\mathscr{A}$-partition coincides with the $\sigma(\mathscr{A})$-partition.

An observer with the information in $\sigma(\mathscr{A})$ knows, not the point ω drawn, but only the equivalence class containing it. That is exactly the information he has. In Example 4.6, it is as though an observer with the items of information in $\mathscr{A}_1$ were unable to combine them to get information about B_{23}.

EXAMPLE 4.9

If $H_n = [\omega\colon d_n(\omega) = 0]$ as in Example 4.5, and if $\mathscr{A} = \{H_1, H_3, H_5, \ldots\}$, then ω and ω' satisfy (4.16) if and only if $d_n(\omega) = d_n(\omega')$ for all odd n. The information in $\sigma(\mathscr{A})$ is thus the set of values of $d_n(\omega)$ for n odd.

One who knows that ω lies in a set A has more information about ω the smaller A is. One who knows $I_A(\omega)$ for each A in a class $\mathscr{A}$, however, has more information about ω the larger $\mathscr{A}$ is. Furthermore, to have the information in $\mathscr{A}_1$ *and* the information in $\mathscr{A}_2$ is to have the information in $\mathscr{A}_1 \cup \mathscr{A}_2$, not that in $\mathscr{A}_1 \cap \mathscr{A}_2$.

The following example points up the informal nature of this interpretation of σ-fields as information.

EXAMPLE 4.10

In the unit interval $(\Omega, \mathscr{F}, P)$ let $\mathscr{G}$ be the σ-field consisting of the countable and the cocountable sets. Since $P(G)$ is 0 or 1 for each G in $\mathscr{G}$ each set H in $\mathscr{F}$ is independent of $\mathscr{G}$. But in this case the $\mathscr{G}$-partition consists of the singletons, and so the information in $\mathscr{G}$ tells the observer exactly which ω in Ω has been drawn. (i) The σ-field $\mathscr{G}$ contains *no* information about H—in the sense that H and $\mathscr{G}$ are independent. (ii) The σ-field $\mathscr{G}$ contains *all* the information about H—in the sense that it tells the observer exactly which ω was drawn.

In this example, (i) and (ii) stand in apparent contradiction. But the mathematics is in (i)—H and $\mathscr{G}$ *are* independent—while (ii) only concerns heuristic interpretation. The source of the difficulty or apparent paradox here lies in the unnatural structure of the σ-field $\mathscr{G}$ rather than in any deficiency in the notion of independence.[†] The heuristic equating of σ-fields and information is helpful even though it sometimes breaks down, and of course proofs are indifferent to whatever illusions and vagaries brought them into existence.

The Borel–Cantelli Lemmas

This is *the first Borel–Cantelli lemma*:

THEOREM 4.3

If $\Sigma_n P(A_n)$ *converges, then* $P(\limsup_n A_n) = 0$.

Proof. From $\limsup_n A_n \subset \bigcup_{k=m}^{\infty} A_k$ follows $P(\limsup_n A_n) \leq P(\bigcup_{k=m}^{\infty} A_k) \leq \Sigma_{k=m}^{\infty} P(A_k)$, and this sum tends to 0 as $m \to \infty$ if $\Sigma_n P(A_n)$ converges. ■

By Theorem 4.1, $P(A_n) \to 0$ implies that $P(\liminf_n A_n) = 0$; in Theorem 4.3 hypothesis and conclusion are both stronger.

EXAMPLE 4.11

Consider the run length $l_n(\omega)$ of Example 4.1 and a sequence $\{r_n\}$ of positive reals. *If the series* $\Sigma 1/2^{r_n}$ *converges, then*

$$P[\omega\colon l_n(\omega) \geq r_n \text{ i.o.}] = 0. \tag{4.17}$$

To prove this, note that if s_n is r_n rounded up to the next integer, then by (4.7), $P[\omega\colon l_n(\omega) \geq r_n] = 2^{-s_n} \leq 2^{-r_n}$. Therefore, (4.17) follows by the first Borel—Cantelli lemma.

[†]See Problem 4.10 for a more extreme example.

If $r_n = (1+\epsilon)\log_2 n$ and ϵ is positive, there is convergence because $2^{-r_n} = 1/n^{1+\epsilon}$. Thus

$$P[\omega\colon\ l_n(\omega) \geq (1+\epsilon)\log_2 n \text{ i.o.}] = 0. \tag{4.18}$$

The limit superior of the ratio $l_n(\omega)/\log_2 n$ exceeds 1 if and only if ω belongs to the set in (4.18) for some positive rational ϵ. Since the union of this countable class of sets has probability 0,

$$P\left[\omega\colon\ \limsup_n \frac{l_n(\omega)}{\log_2 n} > 1\right] = 0. \tag{4.19}$$

To put it the other way around.

$$P\left[\omega\colon\ \limsup_n \frac{l_n(\omega)}{\log_2 n} \leq 1\right] = 1. \tag{4.20}$$

Technically, the probability in (4.20) refers to Lebesgue measure. Intuitively, it refers to an infinite sequence of independent tosses of a fair coin.

In this example, whether $\limsup_n l_n(\omega)/\log_2 n \leq 1$ holds or not is a property of ω, and the property in fact holds for ω in an $\mathscr{F}$-set of probability 1. In such a case the property is said to hold *with probability* 1, or *almost surely*. In nonprobabilistic contexts, a property that holds for ω outside a (measurable) set of measure 0 holds *almost everywhere*, or for *almost all* ω.

EXAMPLE 4.12

The preceding example has an interesting arithmetic consequence. Truncating the dyadic expansion at n gives the standard $(n-1)$-place approximation $\Sigma_{k=1}^{n-1} d_k(\omega)2^{-k}$ to ω; the error is between 0 and 2^{-n+1}, and the error relative to the maximum is

$$e_n(\omega) = \frac{\omega - \Sigma_{k=1}^{n-1} d_k(\omega)2^{-k}}{2^{-n+1}} = \sum_{i=1}^{\infty} d_{n+i-1}(\omega)2^{-i}, \tag{4.21}$$

which lies between 0 and 1. The binary expansion of $e_n(\omega)$ begins with $l_n(\omega)$ 0's, and then comes a 1. Hence $.0\ldots01 \leq e_n(\omega) \leq .0\ldots0111\ldots$, where there are $l_n(\omega)$ 0's in the extreme terms. Therefore,

$$\frac{1}{2^{l_n(\omega)+1}} \leq e_n(\omega) \leq \frac{1}{2^{l_n(\omega)}}, \tag{4.22}$$

so that results on run length give information about the error of approximation.

By the left-hand inequality in (4.22), $e_n(\omega) \le x_n$ (assume that $0 < x_n \le 1$) implies that $l_n(\omega) \ge -\log_2 x_n - 1$; since $\Sigma 2^{-r_n} < \infty$ implies (4.17), $\Sigma x_n < \infty$ implies $P[\omega\colon e_n(\omega) \le x_n \text{ i.o.}] = 0$. (Clearly, $[\omega\colon e_n(\omega) \le x]$ is a Borel set.) In particular,

$$P[\omega\colon e_n(\omega) \le 1/n^{1+\epsilon} \text{ i.o.}] = 0. \tag{4.23}$$

Technically, this probability refers to Lebesgue measure; intuitively, it refers to a point drawn at random from the unit interval.

EXAMPLE 4.13

The final step in the proof of the normal number theorem (Theorem 1.2) was a disguised application of the first Borel–Cantelli lemma. If $A_n = [\omega\colon |n^{-1}s_n(\omega)| \ge n^{-1/8}]$, then $\Sigma P(A_n) < \infty$, as follows by (1.29), and so $P[A_n \text{ i.o.}] = 0$. But for ω in the set complementary to $[A_n \text{ i.o.}]$, $n^{-1}s_n(\omega) \to 0$.

The proof of Theorem 1.6 is also, in effect, an application of the first Borel—Cantelli lemma.

This is *the second Borel–Cantelli lemma*:

THEOREM 4.4

If $\{A_n\}$ is an independent sequence of events and $\Sigma_n P(A_n)$ diverges, then $P(\limsup_n A_n) = 1$.

Proof. It is enough to prove that $P(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c) = 0$ and hence enough to prove that $P(\bigcap_{k=n}^{\infty} A_k^c) = 0$ for all n. Since $1 - x \le e^{-x}$,

$$P\left(\bigcap_{k=n}^{n+j} A_k^c\right) = \prod_{k=n}^{n+j} (1 - P(A_k)) \le \exp\left[-\sum_{k=n}^{n+j} P(A_k)\right].$$

Since $\Sigma_k P(A_k)$ diverges, the last expression tends to 0 as $j \to \infty$, and hence $P(\cap_{k=n}^{\infty} A_k^c) = \lim_j P(\cap_{k=n}^{n+j} A_k^c) = 0$. ■

By Theorem 4.1, $\limsup_n P(A_n) > 0$ implies $P(\limsup_n A_n) > 0$; in Theorem 4.4, the hypothesis $\Sigma_n P(A_n) = \infty$ is weaker but the conclusion is stronger because of the additional hypothesis of independence.

EXAMPLE 4.14

Since the events $[\omega\colon l_n(\omega) = 0] = [\omega\colon d_n(\omega) = 1], n = 1, 2, \ldots$, are independent and have probability $\frac{1}{2}$, $P[\omega\colon l_n(\omega) = 0 \text{ i.o.}] = 1$.

Since the events $A_n = [\omega\colon l_n(\omega) = 1] = [\omega\colon d_n(\omega) = 0, d_{n+1}(\omega) = 1], n = 1, 2, \ldots$, are not independent, this argument is insufficient to prove that

$$P[\omega\colon l_n(\omega) = 1 \text{ i.o.}] = 1. \tag{4.24}$$

But the events $A_2, A_4, A_6, \ldots$ are independent (Theorem 4.2) and their probabilities form a divergent series, and so $P[\omega\colon l_{2n}(\omega) = 1 \text{ i.o.}] = 1$, which implies (4.24).

Significant applications of the second Borel–Cantelli lemma usually require, in order to get around problems of dependence, some device of the kind used in the preceding example.

EXAMPLE 4.15

There is a complement to (4.17): *If r_n is nondecreasing and $\Sigma 2^{-r_n}/r_n$ diverges, then*

$$P[\omega\colon l_n(\omega) \geq r_n \text{ i.o.}] = 1. \tag{4.25}$$

To prove this, note first that if r_n is rounded up to the next integer, then $\Sigma 2^{-r_n}/r_n$ still diverges and (4.25) is unchanged. Assume then that $r_n = r(n)$ is integral, and define $\{n_k\}$ inductively by $n_t = 1$ and $n_{k+1} = n_k + r_{n_k}, k \geq 1$. Let $A_k = [\omega\colon l_{n_k}(\omega) \geq r_{n_k}] = [\omega\colon d_i(\omega) = 0, n_k \leq i < n_{k+1}]$; since the A_k involve nonoverlapping sequences of time indices, it follows by Corollary 2 to Theorem 4.2 that $A_1, A_2, \ldots$ are independent. By the second Borel-Cantelli lemma, $P[A_k \text{ i.o.}] = 1$ if $\Sigma_k P(A_k) = \Sigma_k 2^{-r(n_k)}$ diverges. But since r_n is nondecreasing,

$$\begin{aligned}\sum_{k\geq 1} 2^{-r(n_k)} &= \sum_{k\geq 1} 2^{-r(n_k)} r^{-1}(n_k)(n_{k+1} - n_k) \\ &\geq \sum_{k\geq 1} \sum_{n_k \leq n < n_{k+1}} 2^{-r_n} r_n^{-1} = \sum_{n\geq 1} 2^{-r_n} r_n^{-1}.\end{aligned}$$

Thus the divergence of $\Sigma_n 2^{-r_n} r_n^{-1}$ implies that of $\Sigma_k 2^{-r(n_k)}$, and it follows that, with probability 1, $l_{n_k}(\omega) \geq r_{n_k}$ for infinitely many values of k. But this is stronger than (4.25).

The result in Example 4.2 follows if $r_n \equiv r$, but this is trivial. If $r_n = \log_2 n$, then $\Sigma 2^{-r_n}/r_n = \Sigma 1/(n \log_2 n)$ diverges, and therefore

$$P[\omega\colon l_n(\omega) \geq \log_2 n \text{ i.o.}] = 1. \tag{4.26}$$

By (4.26) and (4.20),

$$P\left[\omega\colon \limsup_n \frac{l_n(\omega)}{\log_2 n} = 1\right] = 1. \tag{4.27}$$

Thus for ω in a set of probability 1, $\log_2 n$ as a function of n is a kind of "upper envelope" for the function $l_n(\omega)$.

EXAMPLE 4.16

By the right-hand inequality in (4.22), if $l_n(\omega) \geq \log_2 n$, then $e_n(\omega) \leq 1/n$. Hence (4.26) gives

$$P\left[\omega\colon e_n(\omega) \leq \frac{1}{n} \text{ i.o.}\right] = 1. \tag{4.28}$$

This and (4.23) show that, with probability 1, $e_n(\omega)$ has $1/n$ as a "lower envelope." The discrepancy between ω and its $(n-1)$-place approximation $\Sigma_{k=1}^{n-1} d_k(\omega)2^{-k}$ will fall infinitely often below $1/(n2^{n-1})$ but not infinitely often below $1/(n^{1+\epsilon}2^{n-1})$.

EXAMPLE 4.17

Examples 4.12 and 4.16 have to do with the approximation of real numbers by rationals: Diophantine approximation. Change the $x_n = 1/n^{1+\epsilon}$ of (4.23) to $1/((n-1)\log 2)^{1+\epsilon}$. Then Σx_n still converges, and hence

$$P[\omega\colon e_n(\omega) \leq 1/(\log 2^{n-1})^{1+\epsilon} \text{ i.o.}] = 0.$$

And by (4.28),

$$P[\omega\colon e_n(\omega) < 1/\log 2^{n-1} \text{ i.o.}] = 1.$$

The dyadic rational $\Sigma_{k=1}^{n-1} d_k(\omega)2^{-k} = p/q$ has denominator $q = 2^{n-1}$, and $e_n(\omega) = q(\omega - p/q)$. There is therefore probability 1 that, if q is restricted to the powers of 2, then $0 \leq \omega - p/q < 1/(q \log q)$ holds for infinitely many p/q but $0 \leq \omega - p/q \leq 1/(q \log^{1+\epsilon} q)$ holds only for finitely many.[†] But contrast this with Theorems 1.5 and 1.6: The sharp rational approximations to a real number come not from truncating its dyadic (or decimal) expansion, but from truncating its continued-fraction expansion; see Section 24.

The Zero–One Law

For a sequence $A_1, A_2, \ldots$ of events in a probability space $(\Omega, \mathscr{F}, P)$ consider the σ-fields $\sigma(A_n, A_{n+1}, \ldots)$ and their intersection

$$\mathscr{T} = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots). \tag{4.29}$$

This is the *tail σ-field* associated with the sequence $\{A_n\}$, and its elements are called *tail events*. The idea is that a tail event is determined solely by the A_n for arbitrarily large n.

[†]This ignores the possibility of even p (reducible p/q); but see Problem 1.11(b). And rounding ω up to $(p+1)/q$ instead of down to p/q changes nothing; see Problem 4.13.

EXAMPLE 4.18
Since $\limsup_m A_m = \cap_{k\geq n} \cup_{i\geq k} A_i$ and $\liminf_m A_m = \cup_{k\geq n} \cap_{i\geq k} A_i$ are both in $\sigma(A_n, A_{n+1}, \ldots)$, the limits superior and inferior are tail events for the sequence $\{A_n\}$.

EXAMPLE 4.19
Let $l_n(\omega)$ be the run length, as before, and let $H_n = [\omega\colon d_n(\omega) = 0]$. For each n_0,

$$[\omega\colon l_n(\omega) \geq r_n \text{ i.o.}] = \bigcap_{n\geq n_0} \bigcup_{k\geq n} [\omega\colon l_k(\omega) \geq r_k]$$
$$= \bigcap_{n\geq n_0} \bigcup_{k\geq n} H_k \cap H_{k+1} \cap \cdots \cap H_{k+r_k-1}.$$

Thus $[\omega\colon l_n(\omega) \geq r_n \text{ i.o.}]$ is a tail event for the sequence $\{H_n\}$.

The probabilities of tail events are governed by *Kolmogorov's zero-one law:*[†]

THEOREM 4.5
If $A_1, A_2, \ldots$ is an independent sequence of events, then for each event A in the tail σ-field (4.29)*, $P(A)$ is either* 0 *or* 1.

Proof. By Corollary 2 to Theorem 4.2, $\sigma(A_1), \ldots, \sigma(A_{n-1}), \sigma(A_n, A_{n+1}, \ldots)$ are independent. If $A \in \mathscr{T}$, then $A \in \sigma(A_n, A_{n+1}, \ldots)$ and therefore $A_1, \ldots, A_{n-1}, A$ are independent. Since independence of a collection of events is defined by independence of each finite subcollection, the sequence $A, A_1, A_2, \ldots$ is independent. By a second application of Corollary 2 to Theorem 4.2, $\sigma(A)$ and $\sigma(A_1, A_2, \ldots)$ are independent. But $A \in \mathscr{T} \subset \sigma(A_1, A_2, \ldots)$; from $A \in \sigma(A)$ and $A \in \sigma(A_1, A_2, \ldots)$ it follows that A is independent of itself: $P(A \cap A) = P(A)P(A)$. This is the same as $P(A) = (P(A))^2$ and can hold only if $P(A)$ is 0 or 1. ■

EXAMPLE 4.20
By the zero–one law and Example 4.18, $P(\limsup_n A_n)$ is 0 or 1 if the A_n are independent. The Borel–Cantelli lemmas in this case go further and give a specific criterion in terms of the convergence or divergence of $\Sigma P(A_n)$.

[†]For a more general version, see Theorem 22.3.

Kolmogorov's result is surprisingly general, and it is in many cases quite easy to use it to show that the probability of some set must have one of the extreme values 0 and 1. It is perhaps curious that it should so often be very difficult to determine which of these extreme values is the right one.

EXAMPLE 4.21

By Kolmogorov's theorem and Example 4.19, $[\omega\colon l_n(\omega) \geq r_n \text{ i.o.}]$ has probability 0 or 1. Call the sequence $\{r_n\}$ an *outer boundary* or an *inner boundary* according as this probability is 0 or 1.

In Example 4.11 it was shown that $\{r_n\}$ is an outer boundary if $\Sigma 2^{-r_n} < \infty$. In Example 4.15 it was shown that $\{r_n\}$ is an inner boundary if r_n is nondecreasing and $\Sigma 2^{-r_n} r_n^{-1} = \infty$. By these criteria $r_n = \theta \log_2 n$ gives an outer boundary if $\theta > 1$ and an inner boundary if $\theta \leq 1$.

What about the sequence $r_n = \log_2 n + \theta \log_2 \log_2 n$? Here $\Sigma 2^{-r_n} = \Sigma 1/n(\log_2 n)^{\theta}$, and this converges for $\theta > 1$, which gives an outer boundary. Now $2^{-r_n} r_n^{-1}$ is of the order $1/n(\log_2 n)^{1+\theta}$, and this diverges if $\theta \leq 0$, which gives an inner boundary (this follows indeed from (4.26)). But this analysis leaves the range $0 < \theta \leq 1$ unresolved, although every sequence is either an inner or an outer boundary. This question is pursued further in Example 6.5.

PROBLEMS

4.1. 2.1 ↑ The limits superior and inferior of a numerical sequence $\{x_n\}$ can be defined as the supremum and infimum of the set of limit points—that is, the set of limits of convergent subsequences. This is the same thing as defining

$$\limsup_n x_n = \bigwedge_{n=1}^{\infty} \bigvee_{k=n}^{\infty} x_k \tag{4.30}$$

and

$$\liminf_n x_n = \bigvee_{n=1}^{\infty} \bigwedge_{k=n}^{\infty} x_k . \tag{4.31}$$

Compare these relations with (4.4) and (4.5) and prove that

$$I_{\limsup_n A_n} = \limsup_n I_{A_n}, \quad I_{\liminf_n A_n} = \liminf_n I_{A_n} .$$

Prove that $\lim_n A_n$ exists in the sense of (4.6) if and only if $\lim_n I_{A_n}(\omega)$ exists for each ω.

4.2. ↑

(a) Prove that

$$(\limsup_n A_n) \cap (\limsup_n B_n) \supset \limsup_n (A_n \cap B_n),$$
$$(\limsup_n A_n) \cup (\limsup_n B_n) = \limsup_n (A_n \cup B_n),$$
$$(\liminf_n A_n) \cap (\liminf_n B_n) = \liminf_n (A_n \cap B_n),$$
$$(\liminf_n A_n) \cup (\liminf_n B_n) \subset \liminf_n (A_n \cup B_n).$$

Show by example that the two inclusions can be strict.

(b) The numerical analogue of the first of the relations in part (a) is

$$(\limsup_n x_n) \wedge (\limsup_n y_n) \geq \limsup_n (x_n \wedge y_n).$$

Write out and verify the numerical analogues of the others.

(c) Show that

$$\limsup_n A_n^c = (\liminf_n A_n)^c,$$
$$\liminf_n A_n^c = (\limsup_n A_n)^c,$$
$$\limsup_n A_n - \liminf_n A_n = \limsup_n (A_n \cap A_{n+1}^c)$$
$$= \limsup_n (A_n^c \cap A_{n+1}).$$

(d) Show that $A_n \to A$ and $B_n \to B$ together imply that $A_n \cup B_n \to A \cup B$ and $A_n \cap B_n \to A \cap B$.

4.3. Let A_n be the square $[(x, y)\colon |x| \leq 1, |y| \leq 1]$ rotated through the angle $2\pi n\theta$. Give geometric descriptions of $\limsup_n A_n$ and $\liminf A_n$ in case

(a) $\theta = \frac{1}{8}$;

(b) θ is rational;

(c) θ is irrational. *Hint*: The $2\pi n\theta$ reduced modulo 2π are dense in $[0, 2\pi]$ if θ is irrational.

(d) When is there convergence is the sense of (4.6)?

4.4. Find a sequence for which all three inequalities in (4.9) are strict.

4.5. **(a)** Show that $\lim_n P(\liminf_k A_n \cap A_k^c) = 0$. *Hint*: Show that $\limsup_n \liminf_k A_n \cap A_k^c$ is empty.

Put $A^* = \limsup_n A_n$ and $A_* = \liminf_n A_n$.

(b) Show that $P(A_n - A^*) \to 0$ and $P(A_* - A_n) \to 0$.

(c) Show that $A_n \to A$ (in the sense that $A = A^* = A_*$) implies $P(A \Delta A_n) \to 0$.

(d) Suppose that A_n converges to A in the weaker sense that $P(A \Delta A^*) = P(A \Delta A_*) = 0$ (which implies that $P(A^* - A_*) = 0$). Show that $P(A \Delta A_n) \to 0$ (which implies that $P(A_n) \to P(A)$).

4.6. In a space of six equally likely points (a die is rolled) find three events that are not independent even though each is independent of the intersection of the other two.

4.7. For events $A_1, \ldots, A_n$, consider the 2^n equations $P(B_1 \cap \cdots \cap B_n) = P(B_1) \cdots P(B_n)$ with $B_i = A_i$ or $B_i = A_i^c$ for each i. Show that $A_1, \ldots, A_n$ are independent if all these equations hold.

4.8. For each of the following classes $\mathscr{A}$, describe the $\mathscr{A}$-partition defined by (4.16).

(a) The class of finite and cofinite sets.

(b) The class of countable and cocountable sets.

(c) A partition (of arbitrary cardinality) of Ω.

(d) The level sets of $\sin x (\Omega = R^1)$.

(e) The σ-field in Problem 3.5.

4.9. 2.9 2.10 ↑ In connection with Example 4.8 and Problem 2.10, prove these facts:

(a) Every set in $\sigma(\mathscr{A})$ is a union of $\mathscr{A}$-equivalence classes.

(b) If $\mathscr{A} = [A_\theta : \theta \in \Theta]$, then the $\mathscr{A}$-equivalence classes have the form $\cap_\theta B_\theta$, where for each θ, B_θ is A_θ or A_θ^c.

(c) Every finite σ-field is generated by a finite partition of Ω.

(d) If $\mathscr{F}_0$ is a field, then each singleton, even each finite set, in $\sigma(\mathscr{F}_0)$ is a countable intersection of $\mathscr{F}_0$-sets.

4.10. 3.2 ↑ There is in the unit interval a set H that is nonmeasurable in the extreme sense that its inner and outer Lebesgue measures are 0 and 1 (see (3.9) and (3.10)): $\lambda_*(H) = 0$ and $\lambda^*(H) = 1$. See Problem 12.4 for the construction.

Let $\Omega = (0, 1]$, let $\mathscr{G}$ consist of the Borel sets in Ω, and let H be the set just described. Show that the class $\mathscr{F}$ of sets of the form $(H \cap G_1) \cup (H^c \cap G_2)$ for G_1 and G_2 in $\mathscr{G}$ is a σ-field and that $P[(H \cap G_1) \cup (H^c \cap G_2)] = \frac{1}{2}\lambda(G_1) + \frac{1}{2}\lambda(G_2)$ consistently defines a probability measure on $\mathscr{F}$. Show that $P(H) = \frac{1}{2}$ and that $P(G) = \lambda(G)$ for $G \in \mathscr{G}$. Show that $\mathscr{G}$ is generated by a countable subclass (see Problem 2.11). Show that $\mathscr{G}$ contains all the singletons and that H and $\mathscr{G}$ are independent.

The construction proves this: *There exist a probability space $(\Omega, \mathscr{F}, P)$, a σ-field $\mathscr{G}$ in $\mathscr{F}$, and a set H in $\mathscr{F}$, such that $P(H) = \frac{1}{2}$, H and $\mathscr{G}$ are independent, and $\mathscr{G}$ is generated by a countable subclass and contains all the singletons.*

Example 4.10 is somewhat similar, but there the σ-field $\mathscr{G}$ is not countably generated and each set in it has probability either 0 or 1. In the present example $\mathscr{G}$ is countably generated and $P(G)$ assumes every value between 0 and 1 as G ranges over $\mathscr{G}$. Example 4.10 is to some extent unnatural because the $\mathscr{G}$ there is not countably generated. The present example, on the other hand, involves the pathological set H. This example is used in Section 33 in connection with conditional probability; see Problem 33.11.

4.11. **(a)** If $A_1, A_2, \ldots$ are independent events, then $P(\cap_{n=1}^{\infty} A_n) = \Pi_{n=1}^{\infty} P(A_n)$ and $P(\cup_{n=1}^{\infty} A_n) = 1 - \Pi_{n=1}^{\infty}(1 - P(A_n))$. Prove these facts and from them derive the second Borel–Cantelli lemma by the well-known relation between infinite series and products.

(b) Show that $P(\limsup_n A_n) = 1$ if for each k the series $\Sigma_{n>k} P(A_n | A_k^c \cap \ldots \cap A_{n-l}^c)$ diverges. From this deduce the second Borel–Cantelli lemma once again.

(c) Show by example that $P(\limsup_n A_n) = 1$ does not follow from the divergence of $\Sigma_n P(A_n | A_1^c \cap \cdots \cap A_{n-1}^c)$ alone.

(d) Show that $P(\limsup_n A_n) = 1$ if and only if $\Sigma_n P(A \cap A_n)$ diverges for each A of positive probability.

(e) If sets A_n are independent and $P(A_n) < 1$ for all n, then $P[A_n \text{ i.o.}] = 1$ if and only if $P(\cup_n A_n) = 1$.

4.12. **(a)** Show (see Example 4.21) that $\log_2 n + \log_2 \log_2 n + \theta \log_2 \log_2 \log_2 n$ is an outer boundary if $\theta > 1$. Generalize.

(b) Show that $\log_2 n + \log_2 \log_2 \log_2 n$ is an inner boundary.

4.13. Let φ be a positive function of integers, and define B_φ as the set of x in (0, 1) such that $|x - p/2^i| < 1/2^i \varphi(2^i)$ holds for infinitely many pairs p, i. Adapting the proof of Theorem 1.6, show directly (without reference to Example 4.12) that $\Sigma_i 1/\varphi(2^i) < \infty$ implies $\lambda(B_\varphi) = 0$.

4.14. 2.19 ↑ Suppose that there are in $(\Omega, \mathscr{F}, P)$ independent events $A_1, A_2, \ldots$ such that, if $\alpha_n = \min\{P(A_n), 1 - P(A_n)\}$, then $\Sigma \alpha_n = \infty$. Show that P is nonatomic.

4.15. 2.18 ↑ Let F be the set of square-free integers—those integers not divisible by any perfect square. Let F_1 be the set of m such that $p^2 | m$ for no $p \le l$, and show that $D(F_1) = \Pi_{p \le l}(1 - p^{-2})$. Show that $P_n(F_1 - F) \le \Sigma_{p>l} p^{-2}$, and conclude that the square-free integers have density $\Pi_p(1 - p^{-2}) = 6/\pi^2$.

4.16. 2.18 ↑ Reconsider Problem 2.18(d). If D were countably additive on $f(\mathscr{M})$, it would extend to $\sigma(\mathscr{M})$. Use the second Borel–Cantelli lemma.

SECTION 5 SIMPLE RANDOM VARIABLES

Definition

Let $(\Omega, \mathscr{F}, P)$ be an arbitrary probability space, and let X be a real-valued function on Ω; X is a *simple random variable* if it has finite range (assumes only finitely many values) and if

$$[\omega\colon X(\omega) = x] \in \mathscr{F} \tag{5.1}$$

for each real x. (Of course, $[\omega\colon X(\omega) = x] = \emptyset \in \mathscr{F}$ for x outside the range of X.) Whether or not X satisfies this condition depends only on $\mathscr{F}$, not on P, but the point of the definition is to ensure that the probabilities $P[\omega\colon X(\omega) = x]$ are defined. Later sections will treat the theory of general random variables, of functions on Ω having arbitrary range; (5.1) will require modification in the general case.

The $d_n(\omega)$ of the preceding section (the digits of the dyadic expansion) are simple random variables on the unit interval: the sets $[\omega\colon d_n(\omega) = 0]$ and $[\omega\colon d_n(\omega) = 1]$ are finite unions of subintervals and hence lie in the σ-field $\mathscr{B}$ of Borel sets in (0, 1]. The Rademacher functions are also simple random variables. Although the concept itself is thus not entirely new, to proceed further in probability requires a systematic theory of random variables and their expected values.

The run lengths $l_n(\omega)$ satisfy (5.1) but are not simple random variables, because they have infinite range (they come under the general theory). In a discrete space, $\mathscr{F}$ consists of all subsets of Ω, so that (5.1) always holds.

It is customary in probability theory to omit the argument ω. Thus X stands for a general value $X(\omega)$ of the function as well as for the function itself, and $[X = x]$ is short for $[\omega\colon X(\omega) = x]$

A finite sum

$$X = \sum_i x_i I_{A_i} \tag{5.2}$$

is a random variable if the A_i form a finite partition of Ω into $\mathscr{F}$-sets. Moreover, every simple random variable can be represented in the form (5.2): for the x_i take the range of X, and put $A_i = [X = x_i]$. But X may have other such representations because $x_i I_{A_i}$ can be replaced by $\Sigma_j x_i I_{A_{ij}}$ if the A_{ij} form a finite decomposition of A_i into $\mathscr{F}$-sets.

If $\mathscr{G}$ is a sub-σ-field of $\mathscr{F}$, a simple random variable X is *measurable* $\mathscr{G}$, or *measurable with respect to* $\mathscr{G}$, if $[X = x] \in \mathscr{G}$ for each x. A simple random variable is by definition always measurable $\mathscr{F}$. Since $[X \in H] = \bigcup [X = x]$, where the union extends over the finitely many x lying both in H and in the range of X, $[X \in H] \in \mathscr{G}$ for every $H \subset R^1$ if X is a simple random variable variable measurable $\mathscr{G}$.

The σ-field $\sigma(X)$ *generated by* X is the smallest σ-field with respect to which X is measurable; that is, $\sigma(X)$ is the intersection of all σ-fields with respect to which X is measurable. For a finite or infinite sequence $X_1, X_2, \ldots$ of simple random variables, $\sigma(X_1, X_2, \ldots)$ is the smallest σ-field with respect to which *each* X_i is measurable. It can be described explicitly in the finite case:

THEOREM 5.1
Let $X_1, \ldots, X_n$ be simple random variables.

(i) *The σ-field $\sigma(X_1, \ldots, X_n)$ consists of the sets*

$$[(X_1, \ldots, X_n) \in H] = [\omega\colon (X_1(\omega), \ldots, X_n(\omega)) \in H] \tag{5.3}$$

for $H \subset R^n$; H in this representation may be taken finite.

(ii) *A simple random variable Y is measurable $\sigma(X_1, \ldots, X_n)$ if and only if*

$$Y = f(X_1, \ldots, X_n) \tag{5.4}$$

for some $f\colon R^n \to R^1$.

Proof. Let $\mathcal{M}$ be the class of sets of the form (5.3). Sets of the form $[(X_1, \ldots, X_n) = (x_1, \ldots, x_n)] = \bigcap_{i=1}^n [X_i = x_i]$ must lie in $\sigma(X_1, \ldots, X_n)$; each set (5.3) is a finite union of sets of this form because $(X_1, \ldots, X_n)$, as a mapping from Ω to R^n, has finite range. Thus $\mathcal{M} \subset \sigma(X_1, \ldots, X_n)$.

On the other hand, $\mathcal{M}$ is a σ-field because $\Omega = [(X_1, \ldots, X_n) \in R^n]$, $[(X_1, \ldots, X_n) \in H]^c = [(X_1, \ldots, X_n) \in H^c]$, and $\cup_j [(X_1, \ldots, X_n) \in H_j] = [(X_1, \ldots, X_n) \cap \bigcup_j H_j]$. But each X_i is measurable with respect to $\mathcal{M}$, because $[X_i = x]$ can be put in the form (5.3) by taking H to consist of those $(x_1, \ldots, x_n)$ in R^n for which $x_i = x$. It follows that $\sigma(X_1, \ldots, X_n)$ is contained in $\mathcal{M}$ and therefore equals $\mathcal{M}$. As intersecting H with the range (finite) of $(X_1, \ldots, X_n)$ in R^n does not affect (5.3), H may be taken finite. This proves (i).

Assume that Y has the form (5.4)—that is, $Y(\omega) = f(X_1(\omega), \ldots, X_n(\omega))$ for every ω. Since $[Y = y]$ can be put in the form (5.3) by taking H to consist of those $x = (x_1, \ldots, x_n)$ for which $f(x) = y$, it follows that Y is measurable $\sigma(X_1, \ldots, X_n)$.

Now assume that Y is measurable $\sigma(X_1, \ldots, X_n)$. Let $y_1, \ldots, y_r$ be the distinct values Y assumes. By part (i), there exist sets $H_1, \ldots, H_r$ in R^n such that

$$[\omega\colon Y(\omega) = y_i] = [\omega\colon (X_1(\omega), \ldots, X_n(\omega)) \in H_i].$$

Take $f = \Sigma_{i=1}^r y_i I_{H_i}$. Although the H_i need not be disjoint, if H_i and H_j share a point of the form $(X_1(\omega), \ldots, X_n(\omega))$, then $Y(\omega) = y_i$ and $Y(\omega) = y_j$, which is impossible if $i \neq j$. Therefore each $(X_1(\omega), \ldots, X_n(\omega))$ lies in exactly one of the H_i, and it follows that $f(X_1(\omega), \ldots, X_n(\omega)) = Y(\omega)$. ■

Since (5.4) implies that Y is measurable $\sigma(X_1, \ldots, X_n)$, it follows in particular that functions of simple random variables are again simple random variables. Thus X^2, e^{tX}, and so on are simple random variables along with X. Taking f to be $\Sigma_{i=1}^n x_i, \Pi_{i=1}^n x_i$, or $\max_{i \le n} x_i$ shows that sums, products, and maxima of simple random variables are simple random variables.

As explained on p. 57, a sub-σ-field corresponds to partial information about ω. From this point of view, $\sigma(X_1, \ldots, X_n)$ corresponds to a knowledge of the values $X_1(\omega), \ldots, X_n(\omega)$. These values suffice to determine the value $Y(\omega)$ if and only if (5.4) holds. The elements of the $\sigma(X_1, \ldots, X_n)$-partition (see (4.16)) are the sets $[X_1 = x_1, \ldots, X_n = x_n]$ for x_i in the range of X_i.

EXAMPLE 5.1

For the dyadic digits $d_n(\omega)$ on the unit interval, d_3 is not measurable $\sigma(d_1, d_2)$; indeed, there exist ω' and ω'' such that $d_1(\omega') = d_1(\omega'')$ and $d_2(\omega') = d_2(\omega'')$ but $d_3(\omega') \ne d_3(\omega'')$, an impossibility if $d_3(\omega) = f(d_1(\omega), d_2(\omega))$ identically in ω. If such an f existed, one could unerringly predict the outcome $d_3(\omega)$ of the third toss from the outcomes $d_1(\omega)$ and $d_2(\omega)$ of the first two.

EXAMPLE 5.2

Let $s_n(\omega) = \Sigma_{k=1}^n r_k(\omega)$ be the partial sums of the Rademacher functions—see (1.14). By Theorem 5.1(ii) s_k is measurable $\sigma(r_1, \ldots, r_n)$ for $k \le n$, and $r_k = s_k - s_{k-1}$ is measurable $\sigma(s_1, \ldots, s_n)$ for $k \le n$. Thus $\sigma(r_1, \ldots, r_n) = \sigma(s_1, \ldots, s_n)$. In random-walk terms, the first n positions contain the same information as the first n distances moved. In gambling terms, to know the gambler's first n fortunes (relative to his initial fortune) is the same thing as to know his gains and losses on each of the first n plays.

EXAMPLE 5.3

An indicator I_A is measurable $\mathscr{G}$ if and only if A lies in $\mathscr{G}$. And $A \in \sigma(A_1, \ldots, A_n)$ if and only if $I_A = f(I_{A_1}, \ldots, I_{A_n})$ for some $f\colon R^n \to R^1$.

Convergence of Random Variables

It is a basic problem, for given random variables X and $X_1, X_2, \ldots$ on a probability space $(\Omega, \mathscr{F}, P)$, to look for the probability of the event that $\lim_n X_n(\omega) = X(\omega)$. The normal number theorem is an example, one where the probability is 1. It is convenient to characterize the complementary event: $X_n(\omega)$ fails to converge to $X(\omega)$ if and only if there is some ϵ such that for no m does $|X_n(\omega) - X(\omega)|$ remain below ϵ for all n exceeding m—that is to say, if and

only if, for some ϵ, $|X_n(\omega) - X(\omega)| \geq \epsilon$ holds for infinitely many values of n. Therefore,

$$[\lim_n X_n = X]^c = \bigcup_\epsilon [|X_n - X| \geq \epsilon \text{ i.o.}], \tag{5.5}$$

where the union can be restricted to rational (positive) ϵ because the set in the union increases as ϵ decreases (compare (2.2)).

The event $[\lim_n X_n = X]$ therefore always lies in the basic σ-field $\mathscr{F}$, and it has probability 1 if and only if

$$P[|X_n - X| \geq \epsilon \text{ i.o.}] = 0 \tag{5.6}$$

for each ϵ (rational or not). The event in (5.6) is the limit superior of the events $[|X_n - X| \geq \epsilon]$, and it follows by Theorem 4.1 that (5.6) implies

$$\lim_n P[|X_n - X| \geq \epsilon] = 0. \tag{5.7}$$

This leads to a definition: If (5.7) holds for each positive ϵ, then X_n is said to *converge to X in probability*, written $X_n \to_P X$.

These arguments prove two facts:

THEOREM 5.2

(i) *There is convergence* $\lim_n X_n = X$ *with probability* 1 *if and only if* (5.6) *holds for each* ϵ.

(ii) *Convergence with probability* 1 *implies convergence in probability.*

Theorem 1.2, the normal number theorem, has to do with the convergence with probability 1 of $n^{-1}\Sigma_{i=1}^n d_i(\omega)$ to $\frac{1}{2}$. Theorem 1.1 has to do instead with the convergence in probability of the same sequence. By Theorem 5.2(ii), then, Theorem 1.1 is a consequence of Theorem 1.2 (see (1.30) and (1.31)). The converse is not true, however—convergence in probability does not imply convergence with probability 1:

EXAMPLE 5.4

Take $X \equiv 0$ and $X_n = I_{A_n}$. Then $X_n \to_P X$ is equivalent to $P(A_n) \to 0$, and $[\lim_n X_n = X]^c = [A_n \text{ i.o.}]$. Any sequence $\{A_n\}$ such that $P(A_n) \to 0$ but $P[A_n \text{ i.o.}] > 0$ therefore gives a counterexample to the converse to Theorem 5.2(ii).

Consider the event $A_n = [\omega\colon\ l_n(\omega) \geq \log_2 n]$ in Example 4.15. Here, $P(A_n) \leq 1/n \to 0$, while $P[A_n \text{ i.o.}] = 1$ by (4.26), and so this is one counterexample. For an example more extreme and more transparent, define

events in the unit interval in the following way. Define the first two by

$$A_1 = (0, \tfrac{1}{2}], \qquad A_2 = (\tfrac{1}{2}, 1].$$

Define the next four by

$$A_3 = (0, \tfrac{1}{4}], \qquad A_4 = (\tfrac{1}{4}, \tfrac{1}{2}], \qquad A_5 = (\tfrac{1}{2}, \tfrac{3}{4}], \qquad A_6 = (\tfrac{3}{4}, 1].$$

Define the next eight, $A_7, \ldots, A_{14}$, as the dyadic intervals of rank 3. And so on. Certainly, $P(A_n) \to 0$, and since each point ω is covered by one set in each successive block of length 2^k, the set $[A_n \text{ i.o.}]$ is all of (0, 1].

Independence

A sequence $X_1, X_2, \ldots$ (finite or infinite) of simple random variables is by definition *independent* if the classes $\sigma(X_1), \sigma(X_2), \ldots$ are independent in the sense of the preceding section. By Theorem 5.1(i), $\sigma(X_i)$ consists of the sets $[X_i \in H]$ for $H \subset R^1$. The condition for independence of $X_1, \ldots, X_n$ is therefore that

$$P[X_1 \in H_1, \ldots, X_n \in H_n] = P[X_1 \in H_1] \cdots P[X_n \in H_n] \tag{5.8}$$

for linear sets $H_1, \ldots, H_n$. The definition (4.10) also requires that (5.8) hold if one or more of the $[X_i \in H_i]$ is suppressed; but taking H_i to be R^1 eliminates it from each side. For an infinite sequence $X_1, X_2, \ldots$, (5.8) must hold for each n. A special case of (5.8) is

$$P[X_1 = x_1, \ldots, X_n = x_n] = P[X_1 = x_1] \cdots P[X_n = x_n]. \tag{5.9}$$

On the other hand, summing (5.9) over $x_1 \in H_1, \ldots, x_n \in H_n$ gives (5.8). Thus the X_i are independent if and only if (5.9) holds for all $x_1, \ldots, x_n$.

Suppose that

$$\begin{matrix} X_{11} & X_{12} & \cdots \\ X_{21} & X_{22} & \cdots \\ \vdots & \vdots & \end{matrix} \tag{5.10}$$

is an independent array of simple random variables. There may be finitely or infinitely many rows, each row finite or infinite. If $\mathscr{A}_i$ consists of the finite intersections $\bigcap_j [X_{ij} \in H_j]$ with $H_j \subset R^1$, an application of Theorem 4.2 shows that the σ-fields $\sigma(X_{i1}, X_{i2}, \ldots), i = 1, 2, \ldots$ are independent. As a consequence, $Y_1, Y_2, \ldots$ are independent if Y_i is measurable $\sigma(X_{i1}, X_{i2}, \ldots)$ for each i.

EXAMPLE 5.5
The dyadic digits $d_1(\omega), d_2(\omega), \ldots$ on the unit interval are an independent sequence of random variables for which

$$P[d_n = 0] = P[d_n = 1] = \tfrac{1}{2}. \tag{5.11}$$

It is because of (5.11) and independence that the d_n give a model for tossing a fair coin.

The sequence $(d_1(\omega), d_2(\omega), \ldots)$ and the point ω determine one another. It can be imagined that ω is determined by the outcomes $d_n(\omega)$ of a sequence of tosses. It can also be imagined that ω is the result of drawing a point at random from the unit interval, and that ω determines the $d_n(\omega)$. In the second interpretation the $d_n(\omega)$ are all determined the instant ω is drawn, and so it should further be imagined that they are then revealed to the coin tosser or gambler one by one. For example, $\sigma(d_1, d_2)$ corresponds to knowing the outcomes of the first two tosses—to knowing not ω but only $d_1(\omega)$ and $d_2(\omega)$—and this does not help in predicting the value $d_3(\omega)$, because $\sigma(d_1, d_2)$ and $\sigma(d_3)$ are independent. See Example 5.1.

EXAMPLE 5.6
Every permutation can be written as a product of cycles. For example,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 5 & 1 & 7 & 4 & 6 & 2 & 3 \end{pmatrix} = (1562)(37)(4).$$

This permutation sends 1 to 5, 2 to 1, 3 to 7, and so on. The cyclic form on the right shows that 1 goes to 5, which goes to 6, which goes to 2, which goes back to 1; and so on. To standardize this cyclic representation, start the first cycle with 1 and each successive cycle with the smallest integer not yet encountered.

Let Ω consist of the $n!$ permutations of $1, 2, \ldots, n$, all equally probable; $\mathscr{F}$ contains all subsets of Ω, and $P(A)$ is the fraction of points in A. Let $X_k(\omega)$ be 1 or 0 according as the element in the kth position in the cyclic representation of the permutation ω completes a cycle or not. Then $S(\omega) = \Sigma_{k=1}^{n} X_k(\omega)$ is the number of cycles in ω. In the example above, $n = 7, X_1 = X_2 = X_3 = X_5 = 0, X_4 = X_6 = X_7 = 1$, and $S = 3$. The following argument shows that $X_1, \ldots, X_n$ are independent and $P[X_k = 1] = 1/(n - k + 1)$. This will lead later on to results on $P[S \in H]$.

The idea is this: $X_1(\omega) = 1$ if and only if the random permutation ω sends 1 to itself, the probability of which is $1/n$. If it happens that $X_1(\omega) = 1$—that ω fixes 1—then the image of 2 is one of $2, \ldots, n$, and $X_2(\omega) = 1$ if and only if this image is in fact 2; the conditional probability of this is $1/(n - 1)$. If $X_1(\omega) = 0$, on the other hand, then ω sends 1 to some $i \neq 1$, so that the image

of i is one of $1, \ldots, i-1, i+1, \ldots, n$, and $X_2(\omega) = 1$ if and only if this image is in fact 1; the conditional probability of this is again $1/(n-1)$. This argument generalizes.

But the details are fussy. Let $Y_1(\omega), \ldots, Y_n(\omega)$ be the integers in the successive positions in the cyclic representation of ω. Fix k, and let A_v be the set where $(X_1, \ldots, X_{k-1}, Y_1, \ldots, Y_k)$ assumes a specific vector of values $v = (x_1, \ldots, x_{k-1}, y_1, \ldots, y_k)$. The A_r form a partition $\mathscr{A}$ of Ω, and if $P[X_k = 1|A_r] = 1/(n-k+1)$ for each v, then by Example 4.7, $P[X_k = 1] = 1/(n-k+1)$ and X_k is independent of $\sigma(\mathscr{A})$ and hence of the smaller σ-field $\sigma(X_1, \ldots, X_{k-1})$. It will follow by induction that $X_1, \ldots, X_n$ are independent.

Let j be the position of the rightmost 1 among $x_1, \ldots, x_{k-1}$ ($j = 0$ if there are none). Then ω lies in A_r if and only if it permutes $y_1, \ldots, y_j$ among themselves (in a way specified by the values $x_1, \ldots, x_{j-1}, x_j = 1, y_1, \ldots, y_j$) and sends each of $y_{j+1}, \ldots, y_{k-1}$ to the y just to its right. Thus A_r contains $(n-k+1)!$ sample points. And $X_k(\omega) = 1$ if and only if ω also sends y_k to y_{j+1}. Thus $A_r \cap [X_k = 1]$ contains $(n-k)!$ sample points, and so the conditional probability of $X_k = 1$ is $1/(n-k+1)$.

Existence of Independent Sequences

The *distribution* of a simple random variable X is the probability measure μ defined for all subsets A of the line by

$$\mu(A) = P[X \in A]. \tag{5.12}$$

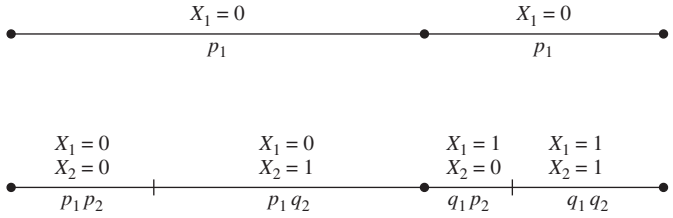
This does define a probability measure. It is discrete in the sense of Example 2.9: If $x_1, \ldots, x_1$ are the distinct points of the range of X, then μ has mass $p_i = P[X = x_i] = \mu\{x_i\}$ at x_i, and $\mu(A) = \Sigma p_i$, the sum extending over those i for which $x_i \in A$. As $\mu(A) = 1$ if A is the range of X, not only is μ discrete, it has finite support.

THEOREM 5.3

Let $\{\mu_n\}$ be a sequence of probability measures on the class of all subsets of the line, each having finite support. There exists on some probability space $(\Omega, \mathscr{F}, P)$ an independent sequence $\{X_n\}$ of simple random variables such that X_n has distribution μ_n.

What matters here is that there are finitely or countably many distributions μ_n. They need not be indexed by the integers; any countable index set will do.

Proof. The probability space will be the unit interval. To understand the construction, consider first the case in which each μ_n concentrates its mass on the two points 0 and 1. Put $p_n = \mu_n\{0\}$ and $q_n = 1 - p_n = \mu_n\{1\}$. Split (0, 1] into two intervals I_0 and I_1 of lengths p_1 and q_1. Define $X_1(\omega) = 0$ for $\omega \in I_0$ and $X_1(\omega) = 1$ for $\omega \in I_1$. If P is Lebesgue measure, then clearly $P[X_1 = 0] = p_1$ and $P[X_1 = 1] = q_1$, so that X_1 has distribution μ_1.



Now split I_0 into two intervals I_{00} and I_{01} of lengths p_1p_2 and p_1q_2, and split I_1 into two intervals I_{10} and I_{11} of lengths q_1p_2 and q_1q_2. Define $X_2(\omega) = 0$ for $\omega \in I_{00} \cup I_{10}$ and $X_2(\omega) = 1$ for $\omega \in I_{01} \cup I_{11}$. As the diagram makes clear, $P[X_1 = 0, X_2 = 0] = p_1p_2$, and similarly for the other three possibilities. It follows that X_1 and X_2 are independent and X_2 has distribution μ_2. Now X_3 is constructed by splitting each of $I_{00}, I_{01}, I_{10}, I_{11}$ in the proportions p_3 and q_3. And so on.

If $p_n = q_n = \frac{1}{2}$ for all n, then the successive decompositions here are the decompositions of (0, 1] into dyadic intervals, and $X_n(\omega) = d_n(\omega)$.

The argument for the general case is not very different. Let $x_{n1}, \ldots, x_{nl_n}$ be the distinct points on which μ_n concentrates its mass, and put $p_{ni} = \mu_n\{x_{ni}\}$ for $1 \le i \le l_n$.

Decompose[†] (0, 1] into l_1 subintervals $I_1^{(1)}, \ldots, I_{l_1}^{(1)}$ of respective lengths $p_{11}, \ldots, p_{1l_1}$. Define X_1 by setting $X_1(\omega) = x_{1i}$ for $\omega \in I_i^{(1)}, 1 \le i \le l_1$. Then ($P$ is Lebesgue measure) $P[\omega\colon X_1(\omega) = x_{1i}] = P(I_i^{(1)}) = p_{1i}, 1 \le i \le l_1$. Thus X_1 is a simple random variable with distribution μ_1.

Next decompose each $I_i^{(1)}$ into l_2 subintervals $I_{i1}^{(2)}, \ldots, I_{il_2}^{(2)}$ of respective lengths $p_{1i}p_{21}, \ldots, p_{1i}p_{2l_2}$. Define $X_2(\omega) = x_{2j}$ for $\omega \in \bigcup_{i=1}^{l_1} I_{ij}^{(2)}, 1 \le j \le l_2$. Then $P[\omega\colon X_1(\omega) = x_{1i}, X_2(\omega) = x_{2j}] = P(I_{ij}^{(2)}) = p_{1i}p_{2j}$. Adding out i shows that $P[\omega\colon X_2(\omega) = x_{2j}] = p_{2j}$, as required. Hence $P[X_1 = x_{1i}, X_2 = x_{2j}] = p_{1i}p_{2j} = P[X_1 = x_{1i}]P[X_2 = x_{2j}]$, and X_1 and X_2 are independent.

The construction proceeds inductively. Suppose that (0, 1] has been decomposed into $l_1 \ldots l_n$ intervals

$$I_{i_1 \ldots i_n}^{(n)}, 1 \le i_1 \le l_1, \ldots, 1 \le i_n \le l_n, \tag{5.13}$$

[†]If $b - a = \delta_1 + \cdots + \delta_l$ and $\delta_i \ge 0$, then $I_i = (a + \Sigma_{j<i}\delta_j, a + \Sigma_{l \le j}\delta_j]$ decomposes $(a, b]$ into subintervals $I_1, \ldots, I_l$ with lengths of δ_i. Of course, I_i is empty if $\delta_i = 0$.

of lengths

$$P(I^{(n)}_{i_1\cdots i_n}) = p_{1,i_1}\cdots p_{n,i_n}. \tag{5.14}$$

Decompose $I^{(n)}_{i_1\cdots i_n}$ into l_{n+1} subintervals $I^{(n+1)}_{i_1\cdots i_n 1},\ldots,I^{(n+1)}_{i_1\cdots i_n l_{n+1}}$ of respective lengths $P(I^{(n)}_{i_1\cdots i_n})p_{n+1,1},\ldots,P(I^{(n)}_{i_1\cdots i_n})p_{n+1,l_{n+1}}$. These are the intervals of the next decomposition. This construction gives a sequence of decompositions (5.13) of (0, 1] into subintervals; each decomposition satisfies (5.14), and each refines the preceding one. If μ_n is given for $1 \le n \le N$, the procedure terminates after N steps; for an infinite sequence it does not terminate at all.

For $1 \le i \le l_n$, put $X_n(\omega) = x_{ni}$ if $\omega \in \bigcup_{i_1\ldots i_{n-1}} I^{(n)}_{i_1\cdots i_{n-1}i}$. Since each decomposition (5.13) refines the preceding, $X_k(\omega) = x_{ki_k}$ for $\omega \in I^{(n)}_{i_1\cdots i_k\cdots i_n}$. Therefore, each element of (5.13) is contained in the element with the same label $i_1\ldots i_n$ in the decomposition

$$A_{i_1\cdots i_n} = [\omega\colon X_1(\omega) = x_{1i_1},\ldots,X_n(\omega) = x_{ni_n}],\ 1 \le i_1 \le l_1,\ldots,1 \le i_n \le l_n.$$

The two decompositions thus coincide, and it follows by (5.14) that $P[X_1 = x_{1i_1},\ldots,X_n = x_{ni_n}] = p_{1,i_1}\cdots p_{n,i_n}$. Adding out the indices $i_1,\ldots,i_{n-1}$ shows that X_n has distribution μ_n and hence that $X_1,\ldots,X_n$ are independent. But n was arbitrary. ■

In the case where the μ_n are all the same, there is an alternative construction based on probabilities in sequence space. Let S be the support (finite) common to the μ_n, and let $p_u, u \in S$, be the probabilities common to the μ_n. In sequence space S^∞, define product measure P on the class $\mathscr{C}_0$ of cylinders by (2.21). By Theorem 2.3, P is countably additive on $\mathscr{C}_0$, and by Theorem 3.1 it extends to $\mathscr{C} = \sigma(\mathscr{C}_0)$. The coordinate functions $z_k(\cdot)$ are random variables on the probability space $(S^\infty, \mathscr{C}, P)$; take these as the X_k. Then (2.22) translates into $P[X_1 = u_1,\ldots,X_n = u_n] = p_{u_1}\cdots p_{u_n}$, which is just what Theorem 5.3 requires in this special case.

Probability theorems such as those in the next sections concern independent sequences $\{X_n\}$ with specified distributions or with distributions having specified properties, and because of Theorem 5.3 these theorems are true not merely in the vacuous sense that their hypotheses are never fulfilled. Similar but more complicated existence theorems will come later. For most purposes the probability space on which the X_n are defined is largely irrelevant. Every independent sequence $\{X_n\}$ satisfying $P[X_n = 1] = p$ and $P[X_n = 0] = 1 - p$ is a model for Bernoulli trials, for example, and for an event like $\bigcup_{n=1}^\infty [\Sigma_{k=1}^n X_k > \alpha n]$, expressed in terms of the X_n alone, the calculation of its probability proceeds in the same way whatever the underlying space $(\Omega, \mathscr{F}, P)$ may be.

It is, of course, an advantage that such results apply not just to some canonical sequence $\{X_n\}$ (such as the one constructed in the proof above) but to every sequence with the appropriate distributions. In some applications of probability within mathematics itself, such as the arithmetic applications of run theory in the preceding section, the underlying Ω does play a role.

Expected Value

A simple random variable in the form (5.2) is assigned *expected value* or *mean value*

$$E[X] = E\left[\sum_i x_i I_{A_i}\right] = \sum_i x_i P(A_i). \tag{5.15}$$

There is the alternative form

$$E[X] = \sum_x xP[X = x], \tag{5.16}$$

the sum extending over the range of X; indeed, (5.15) and (5.16) both coincide with $\Sigma_x \Sigma_{i:x_i=x} x_i P(A_i)$. By (5.16) the definition (5.15) is consistent: different representations (5.2) give the same value to (5.15). From (5.16) it also follows that $E[X]$ depends only on the distribution of X; hence $E[X] = E[Y]$ if $P[X = Y] = 1$.

If X is a simple random variable on the unit interval and if the A_i in (5.2) happen to be subintervals, then (5.15) coincides with the Riemann integral as given by (1.6). More general notions of integral and expected value will be studied later. Simple random variables are easy to work with because the theory of their expected values is transparent and free of technical complications.

As a special case of (5.15) and (5.16),

$$E[I_A] = P(A). \tag{5.17}$$

As another special case, if a constant α is identified with the random variable $X(\omega) \equiv \alpha$, then

$$E[\alpha] = \alpha. \tag{5.18}$$

From (5.2) follows $f(X) = \Sigma_i f(x_i) I_{A_i}$, and hence

$$E[f(X)] = \sum_i f(x_i) P(A_i) = \sum_x f(x) P[X = x], \tag{5.19}$$

the last sum extending over the range of X. For example, the kth *moment* $E[X^k]$ of X is defined by $E[X^k] = \Sigma_y y P[X^k = y]$, where y varies over the range of

X^k, but it is usually simpler to compute it by $E[X^k] = \Sigma_x x^k P[X = x]$, where x varies over the range of X.

If

$$X = \sum_i x_i I_{A_i}, \qquad Y = \sum_j y_j I_{B_j} \tag{5.20}$$

are simple random variables, then $\alpha X + \beta Y = \Sigma_{ij}(\alpha x_i + \beta y_j) I_{A_i \cap B_j}$ has expected value $\Sigma_{ij}(\alpha x_i + \beta y_j) P(A_i \cap B_j) = \alpha \Sigma_i x_i P(A_i) + \beta \Sigma_j y_j P(B_j)$. Expected value is therefore *linear*:

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]. \tag{5.21}$$

If $X(\omega) \le Y(\omega)$ for all ω, then $x_i \le y_j$ if $A_i \cap B_j$ is nonempty, and hence $\Sigma_{ij} x_i P(A_i \cap B_j) \le \Sigma_{ij} y_j P(A_i \cap B_j)$. Expected value therefore *preserves order*:

$$E[X] \le E[Y] \qquad \text{if } X \le Y. \tag{5.22}$$

(It is enough that $X \le Y$ on a set of probability 1.) Two applications of (5.22) give $E[-|X|] \le E[X] \le E[|X|]$, so that by linearity,

$$|E[X]| \le E[|X|]. \tag{5.23}$$

And more generally,

$$|E[X - Y]| \le E[|X - Y|]. \tag{5.24}$$

The relations (5.17) through (5.24) will be used repeatedly, and so will the following theorem on expected values and limits. If there is a finite K such that $|X_n(\omega)| \le K$ for all ω and n, the X_n are *uniformly bounded*.

THEOREM 5.4

If $\{X_n\}$ is uniformly bounded, and if $X = \lim_n X_n$ with probability 1, then $E[X] = \lim_n E[X_n]$.

Proof. By Theorem 5.2(ii), convergence with probability 1 implies convergence in probability: $X_n \to_P X$. And in fact the latter suffices for the present proof. Increase K so that it bounds $|X|$ (which has finite range) as well as all the $|X_n|$; then $|X - X_n| \le 2K$. If $A = [|X - X_n| \ge \epsilon]$, then

$$|X(\omega) - X_n(\omega)| \le 2K I_A(\omega) + \epsilon I_{A^c}(\omega) \le 2K I_A(\omega) + \epsilon$$

for all ω. By (5.17), (5.18), (5.21), and (5.22),

$$E[|X - X_n|] \le 2KP[|X - X_n| \ge \epsilon] + \epsilon.$$

But since $X_n \to_P X$, the first term on the right goes to 0, and since ϵ is arbitrary, $E[|X - X_n|] \to 0$. Now apply (5.24). ■

Theorems of this kind are of constant use in probability and analysis. For the general version, Lebesgue's dominated convergence theorem, see Section 16.

EXAMPLE 5.7
On the unit interval, take $X(\omega)$ identically 0, and take $X_n(\omega)$ to be n^2 if $0 < \omega \le n^{-1}$ and 0 if $n^{-1} < \omega \le 1$. Then $X_n(\omega) \to X(\omega)$ for every ω, although $E[X_n] = n$ does not converge to $E[X] = 0$. Thus Theorem 5.4 fails without some hypothesis such as that of uniform boundedness. See also Example 7.7.

An extension of (5.21) is an immediate consequence of Theorem 5.4:

Corollary. *If $X = \Sigma_n X_n$ on an $\mathscr{F}$-set of probability 1, and if the partial sums of $\Sigma_n X_n$ are uniformly bounded, then $E[X] = \Sigma_n E[X_n]$.*

Expected values for independent random variables satisfy the familiar product law. For X and Y as in (5.20), $XY = \Sigma_{ij} x_i y_j I_{A_i \cap B_j}$. If the x_i are distinct and the y_j are distinct, then $A_i = [X = x_i]$ and $B_j = [Y = y_j]$; for independent X and Y, $P(A_i \cap B_j) = P(A_i)P(B_j)$ by (5.9), and so $E[XY] = \Sigma_{ij} x_i y_j P(A_i)P(B_j) = E[X]E[Y]$. If X, Y, Z are independent, then XY and Z are independent by the argument involving (5.10), so that $E[XYZ] = E[XY]E[Z] = E[X]E[Y]E[Z]$. This obviously extends:

$$E[X_1 \cdots X_n] = E[X_1] \cdots E[X_n] \tag{5.25}$$

if $X_1, \ldots, X_n$ are independent.

Various concepts from discrete probability carry over to simple random variables. If $E[X] = m$, the *variance* of X is

$$\operatorname{Var}[X] = E[(X - m)^2] = E[X^2] - m^2; \tag{5.26}$$

the left-hand equality is a definition, the right-hand one a consequence of expanding the square. Since $\alpha X + \beta$ has mean $\alpha m + \beta$, its variance is $E[((\alpha X + \beta) - (\alpha m + \beta))^2] = E[\alpha^2 (X - m)^2]$:

$$\operatorname{Var}[\alpha X + \beta] = \alpha^2 \operatorname{Var}[X]. \tag{5.27}$$

If $X_1, \ldots, X_n$ have means $m_1, \ldots, m_n$, then $S = \Sigma_{i=1}^n X_i$ has mean $m = \Sigma_{i=1}^n m_i$, and $E[(S - m)^2] = E[(\Sigma_{i=1}^n (X_i - m_i))^2] = \Sigma_{i=1}^n E[(X_i - m_i)^2] + 2\Sigma_{1 \le i < j \le n}$

$E[(X_i - m_i)(X_j - m_j)]$. If the X_i are independent, then so are the $X_i - m_i$, and by (5.25) the last sum vanishes. This gives the familiar formula for the variance of a sum of independent random variables:

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i]. \tag{5.28}$$

Suppose that X is nonnegative; order its range: $0 \le x_1 < x_2 < \cdots < x_k$. Then

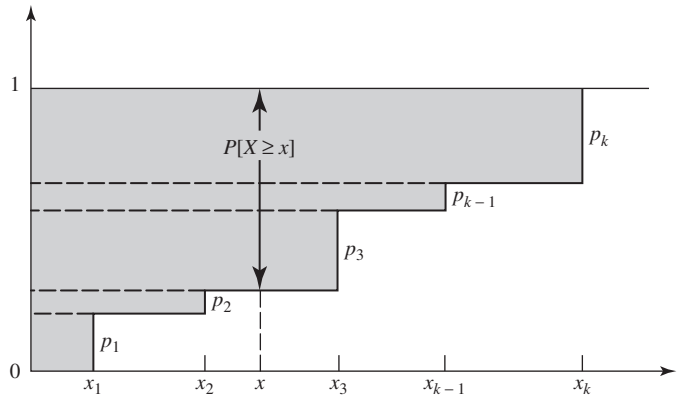
$$\begin{aligned} E[X] &= \sum_{i=1}^{k} x_i P[X = x_i] \\ &= \sum_{i=1}^{k-1} x_i (P[X \ge x_i] - P[X \ge x_{i+1}]) + x_k P[X \ge x_k] \\ &= x_1 P[X \ge x_1] + \sum_{i=2}^{k} (x_i - x_{i-1}) P[X \ge x_i]. \end{aligned}$$

Since $P[X \ge x] = P[X \ge x_1]$ for $0 \le x \le x_1$ and $P[X \ge x] = P[X \ge x_i]$ for $x_{i-1} < x \le x_i$, it is possible to write the final sum as the Riemann integral of a step function:

$$E[X] = \int_0^\infty P[X \ge x]\, dx. \tag{5.29}$$

This holds if X is nonnegative. Since $P[X \ge x] = 0$ for $x > x_k$, the range of integration is really finite.

There is for (5.29) a simple geometric argument involving the "area over the curve." If $p_i = P[X = x_i]$, the area of the shaded region in the figure is the sum $p_1 x_1 + \cdots + p_k x_k = E[X]$ of the areas of the horizontal strips; it is also the integral of the height $P[X \ge x]$ of the region.



Inequalities

There are for expected values several standard inequalities that will be needed. If X is nonnegative, then for positive α (sum over the range of X) $E[X] = \Sigma_x xP[X = x] \geq \Sigma_{x:\ x\geq\alpha} xP[X = x] \geq \alpha\Sigma_{x:\ x\geq\alpha} P[X = x]$. Therefore,

$$P[X \geq \alpha] \leq \frac{1}{\alpha}E[X] \tag{5.30}$$

if X is nonnegative and α positive. A special case of this is (1.20). Applied to $|X|^k$, (5.30) gives *Markov's inequality*,

$$P[|X| \geq \alpha] \leq \frac{1}{\alpha^k}E[|X|^k], \tag{5.31}$$

valid for positive α. If $k = 2$ and $m = E[X]$ is subtracted from X, this becomes the *Chebyshev* (or Chebyshev–Bienaymé) *inequality*:

$$P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2}\text{Var}[X]. \tag{5.32}$$

A function φ on an interval is *convex* [A32] if $\varphi(px + (1-p)y) \leq p\varphi(x) + (1-p)\varphi(y)$ for $0 \leq p \leq 1$ and x and y in the interval. A sufficient condition for this is that φ have a nonnegative second derivative. It follows by induction that $\varphi(\Sigma_{i=1}^{1} p_i x_i) \leq \Sigma_{i=1}^{1} p_i \varphi(x_i)$ if the p_i are nonnegative and add to 1 and the x_i are in the domain of φ. If X assumes the value x_i with probability p_i, this becomes *Jensen's inequality*,

$$\varphi(E[X]) \leq E[\varphi(X)], \tag{5.33}$$

valid if φ is convex on an interval containing the range of X.

Suppose that

$$\frac{1}{p} + \frac{1}{q} = 1, \qquad p > 1,\ q > 1. \tag{5.34}$$

Hölder's inequality is

$$E[|XY|] \leq E^{1/p}[|X|^p] \cdot E^{1/q}[|Y|^q]. \tag{5.35}$$

If, say, the first factor on the right vanishes, then $X = 0$ with probability 1, hence $XY = 0$ with probability 1, and hence the left side vanishes also. Assume then that the right side of (5.35) is positive. If a and b are positive, there exist s and t such that $a = e^{p^{-1}s}$ and $b = e^{q^{-1}t}$. Since e^x is convex, $e^{p^{-1}s+q^{-1}t} \leq p^{-1}e^s + q^{-1}e^t$, or

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

This obviously holds for nonnegative as well as for positive a and b. Let u and v be the two factors on the right in (5.35). For each ω,

$$\left|\frac{X(\omega)Y(\omega)}{uv}\right| \le \frac{1}{p}\left|\frac{X(\omega)}{u}\right|^p + \frac{1}{q}\left|\frac{Y(\omega)}{v}\right|^q$$

Taking expected values and applying (5.34) leads to (5.35).

If $p = q = 2$, Hölder's inequality becomes *Schwarz's inequality*:

$$E[|XY|] \le E^{1/2}[X^2] \cdot E^{1/2}[Y^2]. \tag{5.36}$$

Suppose that $0 < \alpha < \beta$. In (5.35) take $p = \beta/\alpha, q = \beta/(\beta - \alpha)$, and $Y(\omega) = 1$, and replace X by $|X|^\alpha$. The result is *Lyapounov's inequality*,

$$E^{1/\alpha}[|X|^\alpha] \le E^{1/\beta}[|X|^\beta], \quad 0 < \alpha \le \beta. \tag{5.37}$$

PROBLEMS

5.1. **(a)** Show that X is measurable with respect to the σ-field $\mathcal{G}$ if and only if $\sigma(X) \subset \mathcal{G}$. Show that X is measurable $\sigma(Y)$ if and only if $\sigma(X) \subset \sigma(Y)$.

(b) Show that, if $\mathcal{G} = \{\emptyset, \Omega\}$, then X is measurable $\mathcal{G}$ if and only if X is constant.

(c) Suppose that $P(A)$ is 0 or 1 for every A in $\mathcal{G}$. This holds, for example, if $\mathcal{G}$ is the tail field of an independent sequence (Theorem 4.5), or if $\mathcal{G}$ consists of the countable and cocountable sets on the unit interval with Lebesgue measure. Show that if X is measurable $\mathcal{G}$, then $P[X = c] = 1$ for some constant c.

5.2. 2.19 ↑ Show that the unit interval can be replaced by any nonatomic probability measure space in the proof of Theorem 5.3.

5.3. Show that $m = E[X]$ minimizes $E[(X - m)^2]$.

5.4. Suppose that X assumes the values $m - \alpha, m, m + \alpha$ with probabilities $p, 1 - 2p, p$, and show that there is equality in (5.32). Thus Chebyshev's inequality cannot be improved without special assumptions on X.

5.5. Suppose that X has mean m and variance σ^2.

(a) Prove *Cantelli's inequality*

$$P[X - m \ge \alpha] \le \frac{\sigma^2}{\sigma^2 + \alpha^2}, \quad \alpha \ge 0.$$

(b) Show that $P[|X - m| \geq \alpha] \leq 2\sigma^2/(\sigma^2 + \alpha^2)$. When is this better than Chebyshev's inequality?

(c) By considering a random variable assuming two values, show that Cantelli's inequality is sharp.

5.6. The polynomial $E[(t|X| + |Y|)^2]$ in t has at most one real zero. Deduce Schwarz's inequality once more.

5.7. **(a)** Write (5.37) in the form $E^{\beta/\alpha}[|X|^\alpha] \leq E[|X|^\alpha)^{\beta/\alpha}]$ and deduce it directly from Jensen's inequality.

(b) Prove that $E[1/X^p] \geq 1/E^p[X]$ for $p > 0$ and X a positive random variable.

5.8. **(a)** Let f be a convex real function on a convex set C in the plane. Suppose that $(X(\omega), Y(\omega)) \in C$ for all ω and prove a two-dimensional Jensen's inequality:

$$f(E[X], E[Y]) \leq E[f(X, Y)]. \tag{5.38}$$

(b) Show that f is convex if it has continuous second derivatives that satisfy

$$f_{11} \geq 0, \quad f_{22} \geq 0, \quad f_{11} f_{22} \geq f_{12}^2. \tag{5.39}$$

5.9. ↑ Hölder's inequality is equivalent to $E[X^{1/p} Y^{1/q}] \leq E^{1/p}[X] \cdot E^{1/q}[Y]$ $(p^{-1} + q^{-1} = 1)$, where X and Y are nonnegative random variables. Derive this from (5.38).

5.10. ↑ *Minkowski's inequality* is

$$E^{1/p}[|X + Y|^p] \leq E^{1/p}[|X|^p] + E^{1/p}[|Y|^p], \tag{5.40}$$

valid for $p \geq 1$. It is enough to prove that $E[(X^{1/p} + Y^{1/p})^p] \leq (E^{1/p}[X] + E^{1/p}[Y])^p$ for nonnegative X and Y. Use (5.38).

5.11. For events $A_1, A_2, \ldots$, not necessarily independent, let $N_n = \Sigma_{k=1}^n I_{A_k}$ be the number to occur among the first n. Let

$$\alpha_n = \frac{1}{n} \sum_{k=1}^{n} P(A_k), \quad \beta_n = \frac{2}{n(n-1)} \sum_{1 \leq j < k \leq n} P(A_j \cap A_k). \tag{5.41}$$

Show that

$$E[n^{-1} N_n] = \alpha_n, \quad \text{Var}[n^{-1} N_n] = \beta_n - \alpha_n^2 + \frac{\alpha_n - \beta_n}{n}. \tag{5.42}$$

Thus $\text{Var}[n^{-1}N_n] \to 0$ if and only if $\beta_n - \alpha_n^2 \to 0$, which holds if the A_n are independent and $P(A_n) = p$ (Bernoulli trials), because then $\alpha_n = p$ and $\beta_n = p^2 = \alpha_n^2$.

5.12. Show that, if X has nonnegative integers as values, then $E[X] = \Sigma_{n=1}^{\infty} P[X \geq n]$.

5.13. Let $I_i = I_{A_i}$ be the indicators of n events having union A. Let $S_k = \Sigma I_{i_1} \cdots I_{i_k}$, where the summation extends over all k-tuples satisfying $1 \leq i_1 < \cdots < i_k \leq n$. Then $s_k = E[S_k]$ are the terms in the inclusion-exclusion formula $P(A) = s_1 - s_2 + \cdots \pm s_n$. Deduce the inclusion-exclusion formula from $I_A = S_1 - S_2 + \cdots \pm S_n$. Prove the latter formula by expanding the product $\Pi_{i=1}^{n}(1 - I_i)$.

5.14. Let $f_n(x)$ be n^2x or $2n - n^2x$ or 0 according as $0 \leq x \leq n^{-1}$ or $n^{-1} \leq x \leq 2n^{-1}$ or $2n^{-1} \leq x \leq 1$. This gives a standard example of a sequence of continuous functions that converges to 0 but not uniformly. Note that $\int_0^1 f_n(x)dx$ does not converge to 0; relate to Example 5.7.

5.15. By Theorem 5.3, for any prescribed sequence of probabilities p_n, there exists (on some space) an independent sequence of events A_n satisfying $P(A_n) = p_n$. Show that if $p_n \to 0$ but $\Sigma p_n = \infty$, this gives a counterexample (like Example 5.4) to the converse of Theorem 5.2(ii).

5.16. ↑ Suppose that $0 \leq p_n \leq 1$ and put $\alpha_n = \min\{p_n, 1 - p_n\}$. Show that, if $\Sigma\alpha_n$ converges, then on some discrete probability space there exist independent events A_n satisfying $P(A_n) = p_n$. Compare Problem 1.1(b).

5.17. **(a)** Suppose that $X_n \to_p X$ and that f is continuous. Show that $f(X_n) \to_p f(X)$.

(b) Show that $E[|X - X_n|] \to 0$ implies $X_n \to_p X$. Show that the converse is false.

5.18. 2.20 ↑ The proof given for Theorem 5.3 for the special case where the μ_n are all the same can be extended to cover the general case: use Problem 2.20.

5.19. 2.18 ↑ For integers m and primes p, let $\alpha_p(m)$ be the exact power of p in the prime factorization of m: $m = \Pi_p p^{\alpha_p(m)}$. Let $\delta_p(m)$ be 1 or 0 as p divides m or not. Under each P_n (see (2.34)) the α_p and δ_p are random variables. Show that for distinct primes $p_1, \ldots, p_u$,

$$P_n[\alpha_{p_i} \geq k_i, i \leq u] = \frac{1}{n}\left\lfloor \frac{n}{p_1^{k_1} \cdots p_u^{k_u}} \right\rfloor \to \frac{1}{p_1^{k_1} \cdots p_u^{k_u}} \tag{5.43}$$

and

$$P_n[\alpha_{p_i} = k_i, i \leq u] \to \Pi_{i=1}^{u}\left(\frac{1}{p_i^{k_i}} - \frac{1}{p_i^{k_i+1}}\right). \tag{5.44}$$

Similarly,

$$P_n[\delta_{p_i} = 1, i \le u] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor \to \frac{1}{p_1 \cdots p_u}. \tag{5.45}$$

According to (5.44), the α_p are for large n approximately independent under P_n, and according to (5.45), the same is true of the δ_p.

For a function f of positive integers, let

$$E_n[f] = \frac{1}{n}\sum_{m=1}^{n} f(m) \tag{5.46}$$

be its expected value under the probability measure P_n. Show that

$$E_n[\alpha_p] = \sum_{k=1}^{\infty} \frac{1}{n}\left\lfloor \frac{n}{p^k} \right\rfloor \to \frac{1}{p-1}; \tag{5.47}$$

this says roughly that $(p-1)^{-1}$ is the average power of p in the factorization of large integers.

5.20. ↑

(a) From Stirling's formula, deduce

$$E_n[\log] = \log n + O(1). \tag{5.48}$$

From this, the inequality $E_n[\alpha_p] \le 2/p$, and the relation $\log m = \Sigma_p \alpha_p(m) \log p$, conclude that $\Sigma_p P^{-1} \log p$ diverges and that there are infinitely many primes.

(b) Let $\log^* m = \Sigma_p \delta_p(m) \log p$. Show that

$$E_n[\log^*] = \sum_p \frac{1}{n}\left\lfloor \frac{n}{p} \right\rfloor \log p = \log n + O(1). \tag{5.49}$$

(c) Show that $\lfloor 2n/p \rfloor - 2\lfloor n/p \rfloor$ is always nonnegative and equals 1 in the range $n < p \le 2n$. Deduce $E_{2n}[\log^*] - E_n[\log^*] = O(1)$ and conclude that

$$\sum_{p \le x} \log p = O(x). \tag{5.50}$$

Use this to estimate the error removing the integral-part brackets introduces into (5.49), and show that

$$\sum_{p \le x} p^{-1} \log p = \log x + O(1). \tag{5.51}$$

(d) Restrict the range of summation in (5.51) to $\theta x < p \le x$ for an appropriate θ, and conclude that

$$\sum_{p \le x} \log p \asymp x, \tag{5.52}$$

in the sense that the ratio of the two sides is bounded away from 0 and ∞.

(e) Use (5.52) and truncation arguments to prove for the number $\pi(x)$ of primes not exceeding x that

$$\pi(x) \asymp \frac{x}{\log x}. \tag{5.53}$$

(By the prime number theorem the ratio of the two sides in fact goes to 1.) Conclude that the rth prime p_r satisfies $p_r \asymp r \log r$ and that

$$\sum_p \frac{1}{p} = \infty. \tag{5.54}$$

SECTION 6 THE LAW OF LARGE NUMBERS

The Strong Law

Let $X_1, X_2, \ldots$ be a sequence of simple random variables on some probability space $(\Omega, \mathscr{F}, P)$. They are *identically distributed* if their distributions (in the sense of (5.12)) are all the same. Define $S_n = X_1 + \cdots + X_n$. The *strong law of large numbers*:

THEOREM 6.1
If the X_n are independent and identically distributed and $E[X_n] = m$, then

$$P[\lim_n n^{-1} S_n = m] = 1. \tag{6.1}$$

Proof. The conclusion is that $n^{-1}S_n - m = n^{-1}\sum_{i=1}^n (X_i - m) \to 0$ with probability 1. Replacing X_i by $X_i - m$ shows that there is no loss of generality in assuming that $m = 0$. The set in question does lie in $\mathscr{F}$ (see (5.5)), and by Theorem 5.2(i), it is enough to show that $P[|n^{-1}S_n| \ge \epsilon \text{ i.o.}] = 0$ for each ϵ.

Let $E[X_i^2] = \sigma^2$ and $E[X_i^4] = \xi^4$. The proof is like that for Theorem 1.2. First (see (1.26)), $E[S_n^4] = \sum E[X_\alpha X_\beta X_\gamma X_\delta]$, the four indices ranging independently from 1 to n. Since $E[X_i] = 0$, it follows by the product rule (5.25) for independent random variables that the summand vanishes if there is one index different from the three others. This leaves terms of the form $E[X_i^4] = \xi^4$, of

which there are n, and terms of the form $E[X_i^2X_j^2] = E[X_i^2]E[X_j^2] = \sigma^4$ for $i \neq j$, of which there are $3n(n-1)$. Hence

$$E[S_n^4] = n\xi^4 + 3n(n-1)\sigma^4 \leq Kn^2, \tag{6.2}$$

where K does not depend on n.

By Markov's inequality (5.31) for $k = 4, P[|S_n| \geq n\epsilon] \leq Kn^{-2}\epsilon^{-4}$, and so by the first Borel–Cantelli lemma, $P[|n^{-1}S_n| \geq \epsilon \text{ i.o.}] = 0$, as required. ■

EXAMPLE 6.1

The classical example is the strong law of large numbers for Bernoulli trials. Here $P[X_n = 1] = p, P[X_n = 0] = 1 - p, m = p$; S_n represents the number of successes in n trials, and $n^{-1}S_n \to p$ with probability 1. The idea of probability as frequency depends on the long-range stability of the success ratio S_n/n.

EXAMPLE 6.2

Theorem 1.2 is the case of Example 6.1 in which $(\Omega, \mathscr{F}, P)$ is the unit interval and the $X_n(\omega)$ are the digits $d_n(\omega)$ of the dyadic expansion of ω. Here $p = \frac{1}{2}$. The set (1.21) of normal numbers in the unit interval has by (6.1) Lebesgue measure 1; its complement has measure 0 (and so in the terminology of Section 1 is negligible).

The Weak Law

Since convergence with probability 1 implies convergence in probability (Theorem 5.2(ii)), it follows under the hypotheses of Theorem 6.1 that $n^{-1}S_n \to_P m$. But this is of course an immediate consequence of Chebyshev's inequality (5.32) and the rule (5.28) for adding variances:

$$P[|n^{-1}S_n - m| \geq \epsilon] \leq \frac{\text{Var}[S_n]}{n^2\epsilon^2} = \frac{n\,\text{Var}[X_1]}{n^2\epsilon^2} \to 0.$$

This is the *weak law of large numbers*.

Chebyshev's inequality leads to a weak law in other interesting cases as well:

EXAMPLE 6.3

Let Ω_n consist of the $n!$ permutations of $1, 2, \ldots, n$, all equally probable, and let $X_{nk}(\omega)$ be 1 or 0 according as the kth element in the cyclic representation of $\omega \in \Omega_n$ completes a cycle or not. This is Example 5.6, although there the dependence on n was suppressed in the notation. The $X_{n1}, \ldots, X_{nn}$ are independent, and $S_n =$

$X_{n1} + \cdots + X_{nn}$ is the number of cycles. The mean m_{nk} of X_{nk} is the probability that it equals 1, namely $(n-k+1)^{-1}$, and its variance is $\sigma_{nk}^2 = m_{nk}(1-m_{nk})$.

If $L_n = \sum_{k=1}^n k^{-1}$, then S_n has mean $\sum_{k=1}^n m_{nk} = L_n$ and variance $\sum_{k=1}^n m_{nk}(1-m_{nk}) < L_n$. By Chebyshev's inequality,

$$P\left[\left|\frac{S_n - L_n}{L_n}\right| \geq \epsilon\right] < \frac{L_n}{\epsilon^2 L_n^2} = \frac{1}{\epsilon^2 L_n} \to 0.$$

Of the $n!$ permutations on n letters, a proportion exceeding $1 - \epsilon^{-2}L_n^{-1}$ thus have their cycle number in the range $(1 \pm \epsilon)L_n$. Since $L_n = \log n + O(1)$, most permutations on n letters have about $\log n$ cycles. For a refinement, see Example 27.3.

Since Ω_n changes with n, it is the nature of the case that there cannot be a strong law corresponding to this result.

Bernstein's Theorem

Some theorems that can be stated without reference to probability nonetheless have simple probabilistic proofs, as the last example shows. Bernstein's approach to the Weierstrass approximation theorem is another example.

Let f be a function on [0, 1]. The *Bernstein polynomial* of degree n associated with f is

$$B_n(x) = \sum_{k=0}^n f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k} \tag{6.3}$$

THEOREM 6.2
If f is continuous, $B_n(x)$ converges to $f(x)$ uniformly on [0, 1].

According to the Weierstrass approximation theorem, f can be uniformly approximated by polynomials; Bernstein's result goes further and specifies an approximating sequence.

Proof. Let $M = \sup_x |f(x)|$, and let $\delta(\epsilon) = \sup[|f(x) - f(y)|: |x-y| \leq \epsilon]$ be the modulus of continuity of f. It will be shown that

$$\sup_x |f(x) - B_n(x)| \leq \delta(\epsilon) + \frac{2M}{n\epsilon^2}. \tag{6.4}$$

By the uniform continuity of f, $\lim_{\epsilon\to 0}\delta(\epsilon) = 0$, and so this inequality (for $\epsilon = n^{-1/3}$, say) will give the theorem.

Fix $n \geq 1$ and $x \in [0,1]$ for the moment. Let $X_1, \ldots, X_n$ be independent random variables (on some probability space) such that $P[X_i = 1] = x$

and $P[X_i = 0] = 1 - x$; put $S = X_1 + \cdots + X_n$. Since $P[S = k] = \binom{n}{k} x^k (1-x)^{n-k}$, the formula (5.19) for calculating expected values of functions of random variables gives $E[f(S/n)] = B_n(x)$. By the law of large numbers, there should be high probability that S/n is near x and hence (f being continuous) that $f(S/n)$ is near $f(x)$; $E[f(S/n)]$ should therefore be near $f(x)$. This is the probabilistic idea behind the proof and, indeed, behind the definition (6.3) itself.

Bound $|f(n^{-1}S) - f(x)|$ by $\delta(\epsilon)$ on the set $[|n^{-1}S - x| < \epsilon]$ and by $2M$ on the complementary set, and use (5.22) as in the proof of Theorem 5.4. Since $E[S] = nx$, Chebyshev's inequality gives

$$\begin{aligned} |B_n(x) - f(x)| &\le E[|f(n^{-1}S) - f(x)|] \\ &\le \delta(\epsilon) P[|n^{-1}S - x| < \epsilon] + 2MP[|n^{-1}S - x| \ge \epsilon] \\ &\le \delta(\epsilon) + 2M \ \mathrm{Var}[S]/n^2\epsilon^2; \end{aligned}$$

since $\mathrm{Var}[S] = nx(1-x) \le n$, (6.4) follows. ■

A Refinement of the Second Borel–Cantelli Lemma

For a sequence $A_1, A_2, \ldots$ of events, consider the number $N_n = I_{A_1} + \cdots + I_{A_n}$ of occurrences among $A_1, \ldots, A_n$. Since $[A_n \text{ i.o.}] = [\omega\colon \sup_n N_n(\omega) = \infty]$, $P[A_n \text{ i.o.}]$ can be studied by means of the random variables N_n.

Suppose that the A_n are independent. Put $p_k = P(A_k)$ and $m_n = p_1 + \cdots + p_n$. From $E[I_{A_k}] = p_k$ and $\mathrm{Var}[I_{A_k}] = p_k(1-p_k) \le p_k$ follow $E[N_n] = m_n$ and $\mathrm{Var}[N_n] = \sum_{k=1}^n \mathrm{Var}[I_{A_k}] \le m_n$. If $m_n > x$, then

$$\begin{aligned} P[N_n \le x] &\le P[|N_n - m_n| \ge m_n - x] \\ &\le \frac{\mathrm{Var}[N_n]}{(m_n - x)^2} \le \frac{m_n}{(m_n - x)^2}. \end{aligned} \tag{6.5}$$

If $\sum p_n = \infty$, so that $m_n \to \infty$, it follows that $\lim_n P[N_n \le x] = 0$ for each x. Since

$$P[\sup_k N_k \le x] \le P[N_n \le x], \tag{6.6}$$

$P[\sup_k N_k \le x] = 0$ and hence (take the union over $x = 1, 2, \ldots$) $P[\sup_k N_k < \infty] = 0$. Thus $P[A_n \text{ i.o.}] = P[\sup_n N_i = \infty] = 1$ if the A_n are independent and $\sum p_n = \infty$, which proves the second Borel–Cantelli lemma once again.

Independence was used in this argument only to estimate $\mathrm{Var}[N_n]$. Even without independence, $E[N_n] = m_n$ and the first two inequalities in (6.5) hold.

THEOREM 6.3
If $\sum P(A_n)$ *diverges and*

$$\liminf_n \frac{\sum_{j,\ k\le n} P(A_j \cap A_k)}{\left(\sum_{k\le n} P(A_k)\right)^2} \le 1, \tag{6.7}$$

then $P[A_n\ i.o.] = 1$.

As the proof will show, the ratio in (6.7) is at least 1; if (6.7) holds, the inequality must therefore be an equality.

Proof. Let θ_n denote the ratio in (6.7). In the notation above,

$$\begin{aligned}\mathrm{Var}[N_n] &= E[N_n^2] - m_n^2 = \sum_{j,\ k\le n} E[I_{A_j} I_{A_k}] - m_n^2 \\ &= \sum_{j,\ k\le n} P(A_j \cap A_k) - m_n^2 = (\theta_n - 1)m_n^2\end{aligned}$$

(and $\theta_n - 1 \ge 0$). Hence (see (6.5)) $P[N_n \le x] \le (\theta_n - 1)m_n^2/(m_n - x)^2$ for $x < m_n$. Since $m_n^2/(m_n - x)^2 \to 1$, (6.7) implies that $\liminf_n P[N_n \le x] = 0$. It still follows by (6.6) that $P[\sup_k N_k \le x] = 0$, and the rest of the argument is as before. ■

EXAMPLE 6.4
If, as in the second Borel–Cantelli lemma, the A_n are independent (or even if they are merely independent in pairs), the ratio in (6.7) is $1 + \Sigma_{k\le n}(p_k - p_k^2)/m_n^2$, so that $\Sigma P(A_n) = \infty$ implies (6.7).

EXAMPLE 6.5
Return once again to the run lengths $l_n(\omega)$ of Section 4. It was shown in Example 4.21 that $\{r_n\}$ is an outer boundary ($P[l_n \ge r_n \text{ i.o.}] = 0$) if $\Sigma 2^{-r_n} < \infty$. It was also shown that $\{r_n\}$ is an inner boundary ($P[l_n \ge r_n \text{ i.o.}] = 1$) if r_n is nondecreasing and $\Sigma 2^{-r_n} r_n^{-1} = \infty$, but Theorem 6.3 can be used to prove this under the sole assumption that $\Sigma 2^{-r_n} = \infty$.

As usual, the r_n can be taken to be positive integers. Let $A_n = [l_n \ge r_n] = [d_n = \cdots = d_{n+r_n-1} = 0]$. If $j + r_j \le k$, then A_j and A_k are independent. If $j < k < j + r_j$, then $P(A_j|A_k) \le P[d_j = \cdots = d_{k-1} = 0|A_k] = P[d_j = \cdots = d_{k-1} = 0] = 1/2^{k-j}$, and so $P(A_j \cap A_k) \le P(A_k)/2^{k-j}$. Therefore,

$$\sum_{j,\ k\le n} P(A_j \cap A_k)$$

$$\le \sum_{k\le n} P(A_k) + 2\sum_{\substack{j<k\le n\\ j+r_j\le k}} P(A_j)P(A_k) + 2\sum_{\substack{j<k\le n\\ k<j+r_j}} 2^{-(k-j)}P(A_k)$$

$$\le \sum_{k\le n} P(A_k) + \left(\sum_{k\le n} P(A_k)\right)^2 + 2\sum_{k\le n} P(A_k).$$

If $\Sigma P(A_n) = \Sigma 2^{-r_n}$ diverges, then (6.7) follows.

Thus $\{r_n\}$ *is an outer or an inner boundary according as* $\Sigma 2^{-r_n}$ *converges or diverges*, which completely settles the issue. In particular, $r_n = \log_2 n + \theta \log_2 \log_2 n$ gives an outer boundary for $\theta > 1$ and an inner boundary for $\theta \le 1$.

EXAMPLE 6.6

It is now possible to complete the analysis in Examples 4.12 and 4.16 of the relative error $e_n(\omega)$ in the approximation of ω by $\Sigma_{k=1}^{n-1} d_k(\omega)2^{-k}$. If $l_n(\omega) \ge -\log_2 x_n (0 < x_n < 1)$, then $e_n(\omega) \le x_n$ by (4.22). By the preceding example for the case $r_n = -\log_2 x_n, \Sigma x_n = \infty$ implies that $P[\omega\colon e_n(\omega) \le x_n \text{ i.o.}] = 1$. By this and Example 4.12, $[\omega\colon e_n(\omega) \le x_n \text{ i.o.}]$ *has Lebesgue measure* 0 *or* 1 *according as* Σx_n *converges or diverges*.

PROBLEMS

6.1. Show that $Z_n \to Z$ with probability 1 if and only if for every positive ϵ there exists an n such that $P[|Z_k - Z| < \epsilon, n \le k \le m] > 1 - \epsilon$ for all m exceeding n. This describes convergence with probability 1 in "finite" terms.

6.2. Show in Example 6.3 that $P[|S_n - L_n| \ge L_n^{1/2+\epsilon}] \to 0$.

6.3. As in Examples 5.6 and 6.3, let ω be a random permutation of $1, 2, \ldots, n$. Each $k, 1 \le k \le n$, occupies some position in the bottom row of the permutation ω; let $X_{nk}(\omega)$ be the number of smaller elements (between 1 and $k-1$) lying to the right of k in the bottom row. The sum $S_n = X_{n1} + \cdots + X_{nn}$ is the total number of *inversions*—the number of pairs appearing in the bottom row in reverse order of size. For the permutation in Example 5.6 the values of $X_{71}, \ldots, X_{77}$ are 0, 0, 0, 2, 4, 2, 4, and $S_7 = 12$. Show that $X_{n1}, \ldots, X_{nn}$ are independent and $P[X_{nk} = i] = k^{-1}$ for $0 \le i < k$. Calculate $E[S_n]$ and $\text{Var}[S_n]$. Show that S_n is likely to be near $n^2/4$.

6.4. For a function f on [0, 1] write $\|f\| = \sup_x |f(x)|$. Show that, if f has a continuous derivative f', then $\|f - B_n\| \le \epsilon\|f'\| + 2\|f\|/n\epsilon^2$. Conclude that $\|f - B_n\| = O(n^{-1/3})$.

6.5. Prove *Poisson's theorem*: If $A_1, A_2, \ldots$ are independent events, $\overline{p}_n = n^{-1}\Sigma_{i=1}^n P(A_i)$, and $N = \Sigma_{i=1}^n I_{A_i}$, then $n^{-1}N_n - \overline{p}_n \to_P 0$.
In the following problems $S_n = X_1 + \cdots + X_n$.

6.6. Prove *Cantelli's theorem*: If $X_1, X_2, \ldots$ are independent, $E[X_n] = 0$, and $E[X_n^4]$ is bounded, then $n^{-1}S_n \to 0$ with probability 1. The X_n need not be identically distributed.

6.7. **(a)** Let $x_1, x_2, \ldots$ be a sequence of real numbers, and put $s_n = x_1 + \cdots + x_n$. Suppose that $n^{-2}s_{n^2} \to 0$ and that the x_n are bounded, and show that $n^{-1}s_n \to 0$.

(b) Suppose that $n^{-2}S_{n^2} \to 0$ with probability 1 and that the X_n are uniformly bounded ($\sup_{n,\omega}|X_n(\omega)| < \infty$). Show that $n^{-1}S_n \to 0$ with probability 1. Here the X_n need not be identically distributed or even independent.

6.8. ↑ Suppose that $X_1, X_2, \ldots$ are independent and uniformly bounded and $E[X_n] = 0$. Using only the preceding result, the first Borel-Cantelli lemma, and Chebyshev's inequality, prove that $n^{-1}S_n \to 0$ with probability 1.

6.9. ↑ Use the ideas of Problem 6.8 to give a new proof of Borel's normal number theorem, Theorem 1.2. The point is to return to first principles and use only negligibility and the other ideas of Section 1, not the apparatus of Sections 2 through 6; in particular, $P(A)$ is to be taken as defined only if A is a finite, disjoint union of intervals.

6.10. 5.11 6.7 ↑ Suppose that (in the notation of (5.41)) $\beta_n - \alpha_n^2 = O(1/n)$. Show that $n^{-1}N_n - \alpha_n \to 0$ with probability 1. What condition on $\beta_n - \alpha_n^2$ will imply a weak law? Note that independence is not assumed here.

6.11. Suppose that $X_1, X_2, \ldots$ are *m-dependent* in the sense that random variables more than m apart in the sequence are independent. More precisely, let $\mathscr{A}_j^k = \sigma(X_j, \ldots, X_k)$, and assume that $\mathscr{A}_{j_1}^{k_1}, \ldots, \mathscr{A}_{j_l}^{k_l}$ are independent if $k_{i-1} + m < j_i$ for $i = 2, \ldots, l$. (Independent random variables are 0-dependent.) Suppose that the X_n have this property and are uniformly bounded and that $E[X_n] = 0$. Show that $n^{-1}S_n \to 0$. *Hint:* Consider the subsequences $X_1, X_{i+m+1}, X_{i+2(m+1),\ldots}$ for $1 \le i \le m+1$.

6.12. ↑ Suppose that the X_n are independent and assume the values $x_1, \ldots, x_l$ with probabilities $p(x_1), \ldots, p(x_l)$. For $u_1, \ldots, u_k$ a k-tuple of the x_i's, let $N_n(u_1, \ldots, u_k)$ be the frequency of the k-tuple in the first $n + k - 1$ trials, that is, the number of t such that $1 \le t \le n$ and $X_i = u_1, \ldots, X_{i+k-1} = u_k$. Show that with probability 1, all asymptotic relative frequencies are what they should be—that is, with probability 1, $n^{-1}N_n(u_1, \ldots, u_k) \to p(u_1)\cdots p(u_k)$ for every k and every k-tuple $u_1, \ldots, u_k$.

6.13. ↑ A number ω in the unit interval is *completely normal* if, for every base b and every k and every k-tuple of base-b digits, the k-tuple appears in the base-b expansion of ω with asymptotic relative frequency b^{-k}. Show that the set of completely normal numbers has Lebesgue measure 1.

6.14. *Shannon's theorem*. Suppose that $X_1, X_2, \ldots$ are independent, identically distributed random variables taking on the values $1, \ldots, r$ with positive probabilities $p_1, \ldots, p_r$. If $p_n(i_1, \ldots, i_n) = p_{i_1} \ldots p_{i_n}$ and $p_n(\omega) = p_n(X_1(\omega), \ldots, X_n(\omega))$, then $p_n(\omega)$ is the probability that a new sequence of n trials would produce the particular sequence $X_1(\omega), \ldots, X_n(\omega)$ of outcomes that happens actually to have been observed. Show that

$$-\frac{1}{n}\log p_n(\omega) \to h = -\sum_{i=1}^{r} p_i \log p_i$$

with probability 1.

In information theory $1, \ldots, r$ are interpreted as the *letters* of an *alphabet*, $X_1, X_2, \ldots$ are the successive letters produced by an information *source*, and h is the *entropy* of the source. Prove the *asymptotic equipartition property*: For large n there is probability exceeding $1 - \epsilon$ that the probability $p_n(\omega)$ of the observed n-long sequence, or *message*, is in the range $e^{-n(h \pm \epsilon)}$.

6.15. In the terminology of Example 6.5, show that $\log_2 n + \log_2 \log_2 n + \theta \log_2 \log_2 \log_2 n$ is an outer or inner boundary as $\theta > 1$ or $\theta \leq 1$. Generalize. (Compare Problem 4.12.)

6.16. 5.20 ↑ Let $g(m) = \Sigma_p \delta_p(m)$ be the number of distinct prime divisors of m. For $a_n = E_n[g]$ (see (5.46)) show that $a_n \to \infty$. Show that

$$E_n\left[\left(\delta_p - \frac{1}{n}\left\lfloor\frac{n}{p}\right\rfloor\right)\left(\delta_q - \frac{1}{n}\left\lfloor\frac{n}{q}\right\rfloor\right)\right] \leq \frac{1}{np} + \frac{1}{nq} \tag{6.8}$$

for $p \neq q$ and hence that the variance of g under P_n satisfies

$$\mathrm{Var}_n[g] \leq 3 \sum_{p \leq n} \frac{1}{p}. \tag{6.9}$$

Prove the *Hardy-Ramanujan theorem*:

$$\lim_n P_n\left[m\colon \left|\frac{g(m)}{a_n} - 1\right| \geq \epsilon\right] = 0. \tag{6.10}$$

Since $a_n \sim \log\log n$ (see Problem 18.17), most integers under n have something like $\log\log n$ distinct prime divisors. Since $\log\log 10^7$ is a little less than 3, the typical integer under 10^7 has about three prime factors—remarkably few.

6.17. Suppose that $X_1, X_2, \ldots$ are independent and $P[X_n = 0] = p$. Let L_n be the length of the run of 0's starting at the nth place: $L_n = k$ if $X_n = \cdots = X_{n+k-1} = 0 \neq X_{n+k}$. Show that $P[L_n \geq r_n \text{ i.o.}]$ is 0 or 1 according as $\Sigma_n^n p^{r_n}$ converges or diverges. Example 6.5 covers the case $p = \frac{1}{2}$.

SECTION 7 GAMBLING SYSTEMS

Let $X_1, X_2, \ldots$ be an independent sequence of random variables (on some $(\Omega, \mathscr{F}, P)$) taking on the two values $+1$ and -1 with probabilities $P[X_n = +1] = p$ and $P[X_n = -1] = q = 1 - p$. Throughout the section, X_n will be viewed as the gambler's gain on the nth of a series of plays at unit stakes. The game is favorable to the gambler if $p > \frac{1}{2}$, fair if $p = \frac{1}{2}$, and unfavorable if $p < \frac{1}{2}$. The case $p \leq \frac{1}{2}$ will be called the *subfair case*.

After the classical gambler's ruin problem has been solved, it will be shown that every gambling system is in certain respects without effect and that some gambling systems are in other respects optimal. Gambling problems of the sort considered here have inspired many ideas in the mathematical theory of probability, ideas that carry far beyond their origin.

Red-and-black will provide numerical examples. Of the 38 spaces on a roulette wheel, 18 are red, 18 are black, and 2 are green. In betting either on red or on black the chance of winning is $\frac{18}{38}$.

Gambler's Ruin

Suppose that the gambler enters the casino with capital a and adopts the strategy of continuing to bet at unit stakes until his fortune increases to c or his funds are exhausted. What is the probability of ruin, the probability that he will lose his capital, a? What is the probability he will achieve his goal, c? Here a and c are integers.

Let

$$S_n = X_1 + \cdots + X_n, \quad S_0 = 0. \tag{7.1}$$

The gambler's fortune after n plays is $a + S_n$. The event

$$A_{a,\,n} = [a + S_n = c] \cap \bigcap_{k=1}^{n-1} [0 < a + S_k < c] \tag{7.2}$$

represents success for the gambler at time n, and

$$B_{a,\,n} = [a + S_n = 0] \cap \bigcap_{k=1}^{n-1} [0 < a + S_k < c] \tag{7.3}$$

represents ruin at time n. If $s_c(a)$ denotes the probability of ultimate success, then

$$s_c(a) = P(\bigcup_{n=1}^{\infty} A_{a,n}) = \sum_{n=1}^{\infty} P(A_{a,n}) \tag{7.4}$$

for $0 < a < c$.

Fix c and let a vary. For $n \geq 1$ and $0 < a < c$, define $A_{a,n}$ by (7.2), and adopt the conventions $A_{a,0} = \emptyset$ for $0 \leq a < c$ and $A_{c,0} = \Omega$ (success is impossible at time 0 if $a < c$ and certain if $a = c$), as well as $A_{0,n} = A_{c,n} = \emptyset$ for $n \geq 1$ (play never starts if a is 0 or c). By these conventions, $s_c(0) = 0$ and $s_c(c) = 1$.

Because of independence and the fact that the sequence $X_2, X_3, \ldots$ is a probabilistic replica of $X_1, X_2, \ldots$, it seems clear that the chance of success for a gambler with initial fortune a must be the chance of winning the first wager times the chance of success for an initial fortune $a+1$, plus the chance of losing the first wager times the chance of success for an initial fortune $a-1$. It thus seems intuitively clear that

$$s_c(a) = ps_c(a+1) + qs_c(a-1), \quad 0 < a < c. \tag{7.5}$$

For a rigorous argument, define $A'_{a,n}$ just as $A_{a,n}$ but with $S'_n = X_2 + \cdots + X_{n+1}$ in place of S_n in (7.2). Now $P[X_i = x_i, i \leq n] = P[X_{i+1} = x_i, i \leq n]$ for each sequence $x_1, \ldots, x_n$ of $+1$'s and -1's, and therefore $P[(X_1, \ldots, X_n) \in H] = P[(X_2, \ldots, X_{n+1}) \in H]$ for $H \subset R^n$. Take H to be the set of $x = (x_1, \ldots, x_n)$ in R^n satisfying $x_i = \pm 1, a + x_1 + \cdots + x_n = c$, and $0 < a + x_1 + \cdots + x_k < c$ for $k < n$. It follows then that

$$P(A_{a,n}) = P(A'_{a,\,n}). \tag{7.6}$$

Moreover, $A_{a,\,n} = ([X_1 = +1] \cap A'_{a+1,n-1}) \cup ([X_1 = -1] \cap A'_{a-1,n-1})$ for $n \geq 1$ and $0 < a < c$. By independence and (7.6), $P(A_{a,\,n}) = pP(A_{a+1,n-1}) + qP(A_{a-1,n-1})$; adding over n now gives (7.5). Note that this argument involves the entire infinite sequence $X_1, X_2, \ldots$.

It remains to solve the difference equation (7.5) with the side conditions $s_c(0) = 0, s_c(c) = 1$. Let $\rho = q/p$ be the odds against the gambler. Then [A19] there exist constants A and B such that, for $0 \leq a \leq c, s_c(a) = A + B\rho^a$ if $p \neq q$ and $s_c(a) = A + Ba$ if $p = q$. The requirements $s_c(0) = 0$ and $s_c(c) = 1$ determine A and B, which gives the solution:

The probability that the gambler can before ruin attain his goal of c from an initial capital of a is

$$s_c(a) = \begin{cases} \frac{\rho^a - 1}{\rho^c - 1}, & 0 \leq a \leq c, \quad \text{if } \rho = \frac{q}{p} \neq 1, \\ \frac{a}{c}, & 0 \leq a \leq c, \quad \text{if } \rho = \frac{q}{p} = 1. \end{cases} \tag{7.7}$$

EXAMPLE 7.1
The gambler's initial capital is \$900 and his goal is \$1000. If $p = \frac{1}{2}$, his chance of success is very good: $s_{1000}(900) = .9$. At red-and-black, $p = \frac{18}{38}$ and hence $\rho = \frac{20}{18}$; in this case his chance of success as computed by (7.7) is only about .00003.

EXAMPLE 7.2
It is the gambler's desperate intention to convert his \$100 into \$20,000. For a game in which $p = \frac{1}{2}$ (no casino has one), his chance of success is $100/20{,}000 = .005$; at red-and-black it is minute—about 3×10^{-911}.

In the analysis leading to (7.7), replace (7.2) by (7.3). It follows that (7.7) with p and q interchanged (ρ goes to ρ^{-1}) and a and $c-a$ interchanged gives the probability $r_c(a)$ of ruin for the gambler: $r_c(a) = (\rho^{-(c-a)} - 1)/(\rho^{-c} - 1)$ if $\rho \neq 1$ and $r_c(a) = (c-a)/c$ if $\rho = 1$. Hence $s_c(a) + r_c(a) = 1$ holds in all cases: *The probability is* 0 *that play continues forever*.

For positive integers a and b, let

$$H_{a,\,b} = \bigcup_{n=1}^{\infty} \left\{ [S_n = b] \cap \bigcap_{k=1}^{n-1} [-a < S_k < b] \right\}$$

be the event that S_n reaches $+b$ before reaching $-a$. Its probability is simply (7.7) with $c = a + b$: $P(H_{a,\,b}) = s_{a+b}(a)$. Now let

$$H_b = \bigcup_{a=1}^{\infty} H_{a,\,b} = \bigcup_{n=1}^{\infty} [S_n = b] = [\sup_n S_n \geq b]$$

be the event that S_n ever reaches $+b$. Since $H_{a,\,b} \uparrow H_b$ as $a \to \infty$, it follows that $P(H_b) = \lim_a s_{a+b}(a)$; this is 1 if $\rho = 1$ or $\rho < 1$, and it is $1/\rho^b$ if $\rho > 1$. Thus

$$P[\sup_n S_n \geq b] = \begin{cases} 1 & \text{if } p \geq q, \\ (p/q)^b & \text{if } p < q. \end{cases} \tag{7.8}$$

This is the probability that a gambler with unlimited capital can ultimately gain b *units*.

EXAMPLE 7.3
The gambler in Example 7.1 has capital 900 and the goal of winning $b = 100$; in Example 7.2 he has capital 100 and b is 19,900. Suppose, instead, that his capital is infinite. If $p = \frac{1}{2}$, the chance of achieving his goal increases from .9

to 1 in the first example and from .005 to 1 in the second. At red-and-black, however, the two probabilities $.9^{100}$ and $.9^{19900}$ remain essentially what they were before (.00003 and 3×10^{-911}).

Selection Systems

Players often try to improve their luck by betting only when in the preceding trials the wins and losses form an auspicious pattern. Perhaps the gambler bets on the nth trial only when among $X_1, \ldots, X_{n-1}$ there are many more $+1$'s than -1's, the idea being to ride winning streaks (he is "in the vein"). Or he may bet only when there are many more -1's than $+1$'s, the idea being it is then surely time a $+1$ came along (the "maturity of the chances"). There is a mathematical theorem that, translated into gaming language, says all such systems are futile.

It might be argued that it *is* sensible to bet if among $X_1, \ldots, X_{n-1}$ there is an excess of $+1$'s, on the ground that it is evidence of a high value of p. But it is assumed throughout that statistical inference is not at issue: p is fixed—at $\frac{18}{38}$, for example, in the case of red-and-black—and is known to the gambler, or should be.

The gambler's strategy is described by random variables $B_1, B_2, \ldots$ taking the two values 0 and 1: If $B_n = 1$, the gambler places a bet on the nth trial; if $B_n = 0$, he skips that trial. If B_n were $(X_n + 1)/2$, so that $B_n = 1$ for $X_n = +1$ and $B_n = 0$ for $X_n = -1$, the gambler would win every time he bet, but of course such a system requires he be prescient—he must know the outcome X_n in advance. For this reason the value of B_n is assumed to depend only on the values of $X_1, \ldots, X_{n-1}$: there exists some function $b_n\colon R^{n-1} \to R^1$ such that

$$B_n = b_n(X_1, \ldots, X_{n-1}). \tag{7.9}$$

(Here B_1 is constant.) Thus the mathematics avoids, as it must, the question of whether prescience is actually possible.

Define

$$\begin{cases} \mathscr{F}_n = \sigma(X_1, \ldots, X_n), & n = 1, 2, \ldots, \\ \mathscr{F}_0 = \{\emptyset, \Omega\}. \end{cases} \tag{7.10}$$

The σ-field $\mathscr{F}_{n-1}$ generated by $X_1, \ldots, X_{n-1}$ corresponds to a knowledge of the outcomes of the first $n-1$ trials. The requirement (7.9) ensures that B_n is measurable $\mathscr{F}_{n-1}$ (Theorem 5.1) and so depends only on the information actually available to the gambler just before the nth trial.

For $n = 1, 2, \ldots$, let N_n be the time at which the gambler places his nth bet. This nth bet is placed at time k or earlier if and only if the number $\Sigma_{i=1}^{k} B_i$ of

bets placed up to and including time k is n or more; in fact, N_n is the smallest k for which $\Sigma_{i=1}^{k} B_i = n$. Thus the event $[N_n \le k]$ coincides with $[\Sigma_{i=1}^{k} B_i \ge n]$; by (7.9) this latter event lies in $\sigma(B_1, \ldots, B_k) \subset \sigma(X_1, \ldots, X_{k-1}) = \mathscr{F}_{k-1}$. Therefore,

$$[N_n = k] = [N_n \le k] - [N_n \le k-1] \in \mathscr{F}_{k-1}. \tag{7.11}$$

(Even though $[N_n = k]$ lies in $\mathscr{F}_{k-1}$ and hence in $\mathscr{F}$, N_n is, as a function on Ω, generally not a simple random variable, because it has infinite range. This makes no difference, because expected values of the N_n will play no role; (7.11) is the essential property.)

To ensure that play continues forever (stopping rules will be considered later) and that the N_n have finite values with probability 1, make the further assumption that

$$P[B_n = 1 \text{ i.o.}] = 1. \tag{7.12}$$

A sequence $\{B_n\}$ of random variables assuming the values 0 and 1, having the form (7.9), and satisfying (7.12) is a *selection system*.

Let Y_n be the gambler's gain on the nth of the trials at which he does bet: $Y_n = X_{N_n}$. It is only on the set $[B_n = 1 \text{ i.o.}]$ that all the N_n and hence all the Y_n are well defined. To complete the definition, set $Y_n = -1$, say, on $[B_n = 1 \text{ i.o.}]^c$; since this set has probability 0 by (7.12), it really makes no difference how Y_n is defined on it.

Now Y_n is a complicated function on Ω because $Y_n(\omega) = X_{N_n(\omega)}(\omega)$. Nonetheless,

$$[\omega\colon Y_n(\omega) = 1] = \bigcup_{k=1}^{\infty} ([\omega\colon N_n(\omega) = k] \cap [\omega\colon X_k(\omega) = 1])$$

lies in $\mathscr{F}$, and so does its complement $[\omega\colon Y_n(\omega) = -1]$. Hence Y_n is a simple random variable.

EXAMPLE 7.4

An example will fix these ideas. Suppose that the rule is always to bet on the first trail, to bet on the second trial if and only if $X_1 = +1$, to bet on the third trial if and only if $X_1 = X_2$, and to bet on all subsequent trails. Here $B_1 = 1, [B_2 = 1] = [X_1 = +1], [B_3 = 1] = [X_1 = X_2]$, and $B_4 = B_5 = \cdots = 1$. The table shows the ways the gambling can start out. A dot represents a value undetermined by X_1, X_2, X_3. Ignore the rightmost column for the moment.

X_1	X_2	X_3	B_1	B_2	B_3	N_1	N_2	N_3	N_4	Y_1	Y_2	Y_3	τ
-1	-1	-1	1	0	1	1	3	4	5	-1	-1	·	1
-1	-1	$+1$	1	0	1	1	3	4	5	-1	$+1$	·	1
-1	$+1$	-1	1	0	0	1	4	5	6	-1	·	·	1
-1	$+1$	$+1$	1	0	0	1	4	5	6	-1	·	·	1
$+1$	-1	-1	1	1	0	1	2	4	5	$+1$	-1	·	2
$+1$	-1	$+1$	1	1	0	1	2	4	5	$+1$	-1	·	2
$+1$	$+1$	-1	1	1	1	1	2	3	4	$+1$	$+1$	-1	3
$+1$	$+1$	$+1$	1	1	1	1	2	3	4	$+1$	$+1$	$+1$	·

In the evolution represented by the first line of the table, the second bet is placed on the third trial ($N_2 = 3$), which results in a loss because $Y_2 = X_{N_2} = X_3 = -1$. Since $X_3 = -1$, the gambler was "wrong" to bet. But remember that before the third trial he does not know $X_3(\omega)$ (much less ω itself); he knows only $X_1(\omega)$ and $X_2(\omega)$. See the discussion in Example 5.5.

Selection systems achieve nothing because $\{Y_n\}$ has the same structure as $\{X_n\}$:

THEOREM 7.1
For every selection system, $\{Y_n\}$ is independent and $P[Y_n = +1] = p, P[Y_n = -1] = q$.

Proof. Since random variables with indices that are themselves random variables are conceptually confusing at first, the ω's here will not be suppressed as they have been in previous proofs.

Relabel p and q as $p(+1)$ and $p(-1)$, so that $P[\omega\colon X_k(\omega) = x] = p(x)$ for $x = \pm 1$. If $A \in \mathcal{F}_{k-1}$, then A and $[\omega\colon X_k(\omega) = x]$ are independent, and so $P(A \cap [\omega\colon X_k(\omega) = x]) = P(A)p(x)$. Therefore, by (7.11),

$$\begin{aligned} P[\omega\colon Y_n(\omega) = x] &= P[\omega\colon X_{N_n(\omega)}(\omega) = x] \\ &= \sum_{k=1}^{\infty} P[\omega\colon N_n(\omega) = k, X_k(\omega) = x] \\ &= \sum_{k=1}^{\infty} P[\omega\colon N_n(\omega) = k]p(x) \\ &= p(x). \end{aligned}$$

More generally, for any sequence $x_1, \ldots, x_n$ of ± 1's,

$$P[\omega\colon Y_i(\omega) = x_i, i \le n] = P[\omega\colon X_{N_i(\omega)}(\omega) = x_i, i \le n]$$
$$= \sum_{k_1 < \cdots < k_n} P[\omega\colon N_i(\omega) = k_i, X_{k_i}(\omega) = x_i, i \le n],$$

where the sum extends over n-tuples of positive integers satisfying $k_1 < \cdots < k_n$. The event $[\omega\colon N_i(\omega) = k_i, i \le n] \cap [\omega\colon X_{k_i}(\omega) = x_i, i < n]$ lies in $\mathscr{T}_{k_n - 1}$ (note that there is no condition on $X_{k_n}(\omega)$), and therefore

$$P[\omega\colon Y_i(\omega) = x_i, i \le n]$$
$$= \sum_{k_1 < \cdots < k_n} P([\omega\colon N_i(\omega) = k_i, i \le n]$$
$$\cap [\omega\colon X_{k_i}(\omega) = x_i, i < n])p(x_n).$$

Summing k_n over $k_{n-1} + 1, k_{n-1} + 2, \ldots$ brings this last sum to

$$\sum_{k_1 < \cdots < k_{n-1}} P[\omega\colon N_i(\omega) = k_i, X_{k_i}(\omega) = x_i, i < n]p(x_n)$$
$$= P[\omega\colon X_{N_i(\omega)}(\omega) = x_i, i < n]\, p(x_n)$$
$$= P[\omega\colon Y_i(\omega) = x_i, i < n]p(x_n).$$

It follows by induction that

$$P[\omega\colon Y_i(\omega) = x_i, i \le n] = \prod_{i \le n} p(x_i) = \prod_{i \le n} P[\omega\colon Y_i(\omega) = x_i],$$

and so the Y_i are independent (see (5.9)). ■

Gambling Policies

There are schemes that go beyond selection systems and tell the gambler not only whether to bet but how much. Gamblers frequently contrive or adopt such schemes in the confident expectation that they can, by pure force of arithmetic, counter the most adverse workings of chance. If the wager specified for the nth trial is in the amount W_n and the gambler cannot see into the future, then W_n must depend only on $X_1, \ldots, X_{n-1}$. Assume therefore that W_n is a nonnegative function of these random variables: there is an $f_n\colon R^{n-1} \to R^1$ such that

$$W_n = f_n(X_1, \ldots, X_{n-1}) \ge 0. \tag{7.13}$$

Apart from nonnegativity there are at the outset no constraints on the f_n, although in an actual casino their values must be integral multiples of a basic unit. Such a

sequence $\{W_n\}$ is a *betting system*. Since $W_n = 0$ corresponds to a decision not to bet at all, betting systems in effect include selection systems. In the double-or-nothing system, $W_n = 2^{n-1}$ if $X_1 = \cdots = X_{n-1} = -1 (W_1 = 1)$ and $W_n = 0$ otherwise.

The amount the gambler wins on the nth play is $W_n X_n$. If his fortune at time n is F_n, then

$$F_n = F_{n-1} + W_n X_n. \tag{7.14}$$

This also holds for $n = 1$ if F_0 is taken as his initial (nonrandom) fortune. It is convenient to let W_n depend on F_0 as well as the past history of play and hence to generalize (7.13) to

$$W_n = g_n(F_0, X_1, \ldots, X_{n-1}) \geq 0 \tag{7.15}$$

for a function g_n: $R^n \to R^1$. In expanded notation, $W_n(\omega) = g_n(F_0, X_1(\omega), \ldots, X_{n-1}(\omega))$. The symbol W_n does not show the dependence on ω or on F_0, either. For each *fixed* initial fortune F_0, W_n is a simple random variable; by (7.15) it is measurable $\mathscr{F}_{n-1}$. Similarly, F_n is a function of F_0 as well as of $X_1(\omega), \ldots, X_n(\omega)$: $F_n = F_n(F_0, \omega)$.

If $F_0 = 0$ and $g_n \equiv 1$, the F_n reduce to the partial sums (7.1).

Since $\mathscr{F}_{n-1}$ and $\sigma(X_n)$ are independent, and since W_n is measurable $\mathscr{F}_{n-1}$ (for each fixed F_0), W_n and X_n are independent. Therefore, $E[W_n X_n] = E[W_n] \cdot E[X_n]$. Now $E[X_n] = p - q \leq 0$ in the subfair case ($p \leq \frac{1}{2}$), with equality in the fair case ($p = \frac{1}{2}$). Since $E[W_n] \geq 0$, (7.14) implies that $E[F_n] \leq E[F_{n-1}]$. Therefore,

$$F_0 \geq E[F_1] \geq \cdots \geq E[F_n] \geq \cdots \tag{7.16}$$

in the subfair case, and

$$F_0 = E[F_1] = \cdots = E[F_n] = \cdots \tag{7.17}$$

in the fair case. (If $p < q$ and $P[W_n > 0] > 0$, there is strict inequality in (7.16).) Thus no betting system can convert a subfair game into a profitable enterprise.

Suppose that in addition to a betting system, the gambler adopts some policy for quitting. Perhaps he stops when his fortune reaches a set target, or his funds are exhausted, or the auguries are in some way dissuasive. The decision to stop must depend only on the initial fortune and the history of play up to the present.

Let $\tau(F_0, \omega)$ be a nonnegative integer for each ω in Ω and each $F_0 \geq 0$. If $\tau = n$, the gambler plays on the nth trial (betting W_n) and then stops; if $\tau = 0$, he does not begin gambling in the first place. The event $[\omega\colon \tau(F_0, \omega) = n]$

represents the decision to stop just after the nth trial, and so, whatever value F_0 may have, it must depend only on $X_1, \ldots, X_n$. Therefore, assume that

$$[\omega\colon \tau(F_0, \omega) = n] \in \mathscr{F}_n, \qquad n = 0, 1, 2, \ldots \tag{7.18}$$

A τ satisfying this requirement is a *stopping time*. (In general it has infinite range and hence is not a simple random variable; as expected values of τ play no role here, this does not matter.) It is technically necessary to let $\tau(F_0, \omega)$ be undefined or infinite on an ω-set of probability 0. This has no effect on the requirement (7.18), which must hold for each finite n. But it *is* assumed that τ *is finite with probability* 1: play is certain to terminate.

A betting system together with a stopping time is a *gambling policy*. Let π denote such a policy.

EXAMPLE 7.5

Suppose that the betting system is given by $W_n = B_n$, with B_n as in Example 7.4. Suppose that the stopping rule is to quit after the first loss of a wager. Then $[\tau = n] = \cup_{k=1}^{n}[N_k = n, Y_1 = \cdots = Y_{k-1} = +1, Y_k = -1]$. For $j \le k \le n$, $[N_k = n, Y_j = x] = \cup_{m=1}^{n}[N_k = n, N_j = m, X_m = x]$ lies in $\mathscr{F}_n$ by (7.11); hence τ is a stopping time. The values of τ are shown in the rightmost column of the table.

The sequence of fortunes is governed by (7.14) until play terminates, and then the fortune remains for all future time fixed at F_τ (with value $F_{\tau(F_0,\omega)}(\omega)$). Therefore, the gambler's fortune at time n is

$$F_n^* = \begin{cases} F_n & \text{if } \tau \ge n, \\ F_\tau & \text{if } \tau \le n. \end{cases} \tag{7.19}$$

Note that the case $\tau = n$ is covered by both clauses here. If $n - 1 < n \le \tau$, then $F_n^* = F_n = F_{n-1} + W_n X_n = F_{n-1}^* + W_n X_n$; if $\tau \le n - 1 < n$, then $F_n^* = F_\tau = F_{n-1}^*$. Therefore, if $W_n^* = I_{[\tau \ge n]} W_n$, then

$$F_n^* = F_{n-1}^* + I_{[\tau \ge n]} W_n X_n = F_{n-1}^* + W_n^* X_n. \tag{7.20}$$

But this is the equation for a new betting system in which the wager placed at time n is W_n^*. If $\tau \ge n$ (play has not already terminated), W_n^* is the old amount W_n; if $\tau < n$ (play has terminated), W_n^* is 0. Now by (7.18), $[\tau \ge n] = [\tau < n]^c$ lies in $\mathscr{F}_{n-1}$. Thus $I_{[\tau \ge n]}$ is measurable $\mathscr{F}_{n-1}$, so that W_n^* as well as W_n is measurable $\mathscr{F}_{n-1}$, and $\{W_n^*\}$ represents a legitimate betting system. Therefore, (7.16) and (7.17) apply to the new system:

$$F_0 = F_0^* \ge E[F_1^*] \ge \cdots \ge E[F_n^*] \ge \cdots \tag{7.21}$$

if $p \le \frac{1}{2}$, and

$$F_0 = F_0^* = E[F_1^*] = \cdots = E[F_n^*] = \ldots \tag{7.22}$$

if $p = \frac{1}{2}$.

The gambler's ultimate fortune is F_τ. Now $\lim_n F_n^* = F_\tau$ with probability 1, since in fact $F_n^* = F_\tau$ for $n \ge \tau$. If

$$\lim_n E[F_n^*] = E[F_\tau], \tag{7.23}$$

then (7.21) and (7.22), respectively, imply that $E[F_\tau] \le F_0$ and $E[F_\tau] = F_0$. According to Theorem 5.4, (7.23) does hold if the F_n^* are uniformly bounded.

Call the policy *bounded by* M (M nonrandom) if

$$0 \le F_n^* \le M, \qquad n = 0, 1, 2, \ldots. \tag{7.24}$$

If F_n^* is not bounded above, the gambler's adversary must have infinite capital. A negative F_n^* represents a debt, and if F_n^* is not bounded below, the gambler must have a patron of infinite wealth and generosity from whom to borrow and so must in effect have infinite capital. In case F_n^* is bounded below, 0 is the convenient lower bound—the gambler is assumed to have in hand all the capital to which he has access. In any real case, (7.24) holds and (7.23) follows. (There is a technical point that arises because the general theory of integration has been postponed: F_τ must be assumed to have finite range so that it will be a simple random variable and hence have an expected value in the sense of Section 5.[†]) The argument has led to this result:

THEOREM 7.2

For every policy, (7.21) *holds if* $p \le \frac{1}{2}$ *and* (7.22) *holds if* $p = \frac{1}{2}$. *If the policy is bounded* (*and* F_τ *has finite range*), *then* $E[F_\tau] \le F_0$ *for* $p \le \frac{1}{2}$ *and* $E[F_\tau] = F_0$ *for* $p = \frac{1}{2}$.

EXAMPLE 7.6

The gambler has initial capital a and plays at unit stakes until his capital increases to $c (0 \le a \le c)$ or he is ruined. Here $F_0 = a$ and $W_n = 1$, and so $F_n = a + S_n$. The policy is bounded by c, and F_τ is c or 0 according as the gambler succeeds or fails. If $p = \frac{1}{2}$ and if s is the probability of success, then $a = F_0 = E[F_\tau] = sc$. Thus $s = a/c$. This gives a new derivation of (7.7) for the case $p = \frac{1}{2}$. The argument assumes however that play is certain to terminate. If $p \le \frac{1}{2}$, Theorem 7.2 only gives $s \le a/c$, which is weaker than (7.7).

[†]See Problem 7.11.

EXAMPLE 7.7
Suppose as before that $F_0 = a$ and $W_n = 1$, so that $F_n = a + S_n$, but suppose the stopping rule is to quit as soon as F_n reaches $a+b$. Here F_n^* is bounded above by $a+b$ but is not bounded below. If $p = \frac{1}{2}$, the gambler is by (7.8) certain to achieve his goal, so that $F_\tau = a + b$. In this case $F_0 = a < a + b = E[F_\tau]$. This illustrates the effect of infinite capital. It also illustrates the need for uniform boundedness in Theorem 5.4 (compare Example 5.7).

For some other systems (gamblers call them "martingales"), see the problems. For most such systems there is a large chance of a small gain and a small chance of a large loss.

Bold Play†

The formula (7.7) gives the chance that a gambler betting unit stakes can increase his fortune from a to c before being ruined. Suppose that a and c happen to be even and that at each trial the wager is two units instead of one. Since this has the effect of halving a and c, the chance of success is now

$$\frac{\rho^{a/2} - 1}{\rho^{c/2} - 1} = \frac{\rho^a - 1}{\rho^c - 1}\frac{\rho^{c/2} + 1}{\rho^{a/2} + 1}, \quad \frac{q}{p} = \rho \neq 1.$$

If $\rho > 1$ $\left(p < \frac{1}{2}\right)$, the second factor on the right exceeds 1: Doubling the stakes increases the probability of success in the unfavorable case $\rho > 1$. In the case $\rho = 1$, the probability remains the same.

There is a sense in which large stakes are optimal. It will be convenient to rescale so that the initial fortune satisfies $0 \le F_0 \le 1$ and the goal is 1. The policy of *bold play* is this: At each stage the gambler bets his entire fortune, unless a win would carry him past his goal of 1, in which case he bets just enough that a win would exactly achieve that goal:

$$W_n = \begin{cases} F_{n-1} & \text{if } 0 \le F_{n-1} \le \frac{1}{2}, \\ 1 - F_{n-1} & \text{if } \frac{1}{2} \le F_{n-1} \le 1. \end{cases} \tag{7.25}$$

(It is convenient to allow even irrational fortunes.) As for stopping, the policy is to quit as soon as F_n reaches 0 or 1.

Suppose that play has not terminated by time $k-1$; under the policy (7.25), if play is not to terminate at time k, then X_k must be $+1$ or -1 according as $F_{k-1} \le \frac{1}{2}$ or $F_{k-1} \ge \frac{1}{2}$, and the conditional probability of this is at most $m = \max\{p, q\}$. It follows by induction that the probability that bold play continues

†This topic may be omitted.

beyond time n is at most m^n, and so play is certain to terminate (τ is finite with probability 1).

It will be shown that in the subfair case, bold play maximizes the probability of successfully reaching the goal of 1. This is the *Dubins–Savage theorem*. It will further be shown that there are other policies that are also optimal in this sense, and this maximum probability will be calculated. Bold play can be substantially better than betting at constant stakes. This contrasts with Theorems 7.1 and 7.2 concerning respects in which gambling systems are worthless.

From now on, consider only policies π that are bounded by 1 (see (7.24)). Suppose further that play stops as soon as F_n reaches 0 or 1 and that this is certain eventually to happen. Since F_τ assumes the values 0 and 1, and since $[F_\tau = x] = \bigcup_{n=0}^{\infty}[\tau = n] \cap [F_n = x]$ for $x = 0$ and $x = 1$, F_τ is a simple random variable. Bold play is one such policy π.

The policy π leads to success if $F_\tau = 1$. Let $Q_\pi(x)$ be the probability of this for an initial fortune $F_0 = x$:

$$Q_\pi(x) = P[F_\tau = 1] \quad \text{for } F_0 = x. \tag{7.26}$$

Since F_n is a function $\psi_n(F_0, X_1(\omega), \ldots, X_n(\omega)) = \Psi_n(F_0, \omega)$, (7.26) in expanded notation is $Q_\pi(x) = P[\omega\colon\ \Psi_{\tau(x,\omega)}(x, \omega) = 1]$. As π specifies that play stops at the boundaries 0 and 1,

$$\begin{aligned} &Q_\pi(0) = 0, \quad Q_\pi(1) = 1, \\ &0 \le Q_\pi(x) \le 1, \qquad 0 \le x \le 1. \end{aligned} \tag{7.27}$$

Let Q be the Q_π for bold play. (The notation does not show the dependence of Q and Q_π on p, which is fixed.)

THEOREM 7.3
In the subfair case, $Q_\pi(x) \le Q(x)$ for all π and all x.

Proof. Under the assumption $p \le q$, it will be shown later that

$$Q(x) \ge pQ(x + t) + qQ(x - t), \quad 0 \le x - t \le x \le x + t \le 1. \tag{7.28}$$

This can be interpreted as saying that the chance of success under bold play starting at x is at least as great as the chance of success if the amount t is wagered and bold play then pursued from $x+t$ in case of a win and from $x-t$ in case of a loss. Under the assumption of (7.28), optimality can be proved as follows.

Consider a policy π, and let F_n and F_n^* be the simple random variables defined by (7.14) and (7.19) *for this policy*. Now $Q(x)$ is a real function, and so $Q(F_n^*)$ is also a simple random variable; it can be interpreted as the conditional

chance of success if π is replaced by bold play after time n. By (7.20), $F_n^* = x + tX_n$ if $F_{n-1}^* = x$ and $W_n^* = t$. Therefore,

$$Q(F_n^*) = \sum_{x,t} I_{[F_{n-1}^*=x, W_n^*=t]} Q(x + tX_n),$$

where x and t vary over the (finite) ranges of F_{n-1}^* and W_n^*, respectively.

For each x and t, the indicator above is measurable $\mathscr{F}_{n-1}$ and $Q(x + tX_n)$ is measurable $\sigma(X_n)$; since the X_n are independent, (5.25) and (5.17) give

$$E[Q(F_n^*)] = \sum_{x,\ t} P[F_{n-1}^* = x, W_n^* = t] E[Q(x + tX_n)] \tag{7.29}$$

By (7.28), $E[Q(x + tX_n)] \le Q(x)$ if $0 \le x - t \le x \le x + t \le 1$. As it is assumed of π that F_n^* lies in [0, 1] (that is, $W_n^* \le \min\{F_{n-1}^*, 1 - F_{n-1}^*\}$), the probability in (7.29) is 0 unless x and t satisfy this constraint. Therefore,

$$\begin{aligned} E[Q(F_n^*)] &\le \sum_{x,\ t} P[F_{n-1}^* = x, W_n^* = t] Q(x) \\ &= \sum_{x} P[F_{n-1}^* = x] Q(x) = E[Q(F_{n-1}^*)]. \end{aligned}$$

This is true for each n, and so $E[Q(F_n^*)] \le E[Q(F_0^*)] = Q(F_0)$. Since $Q(F_n^*) = Q(F_\tau)$ for $n \ge \tau$, Theorem 5.4 implies that $E[Q(F_\tau)] \le Q(F_0)$. Since $x = 1$ implies that $Q(x) = 1, P[F_\tau = 1] \le E[Q(F_\tau)] \le Q(F_0)$. Thus $Q_\pi(F_0) \le Q(F_0)$ for the policy π, whatever F_0 may be.

It remains to analyze Q and prove (7.28). Everything hinges on the functional equation

$$Q(x) = \begin{cases} pQ(2x), & 0 \le x \le \frac{1}{2}, \\ p + qQ(2x - 1), & \frac{1}{2} \le x \le 1. \end{cases} \tag{7.30}$$

For $x = 0$ and $x = 1$ this is obvious because $Q(0) = 0$ and $Q(1) = 1$. The idea is this: Suppose that the initial fortune is x. If $x \le \frac{1}{2}$, the first stake under bold play is x; if the gambler is to succeed in reaching 1, he must win the first trial (probability p) and then from his new fortune $x + x = 2x$ go on to succeed (probability $Q(2x)$); this makes the first half of (7.30) plausible. If $x \ge \frac{1}{2}$, the first stake is $1 - x$; the gambler can succeed either by winning the first trial (probability p) or by losing the first trial (probability q) and then going on from his new fortune $x - (1 - x) = 2x - 1$ to succeed (probability $Q(2x - 1)$); this makes the second half of (7.30) plausible.

It is also intuitively clear that $Q(x)$ must be an increasing function of x ($0 \le x \le 1$): the more money the gambler starts with, the better off he is. Finally, it is intuitively clear that $Q(x)$ ought to be a continuous function of the initial fortune x.

A formal proof of (7.30) can be constructed as for the difference equation (7.5). If $\beta(x)$ is x for $x \le \frac{1}{2}$ and $1-x$ for $x \ge \frac{1}{2}$, then under bold play $W_n = \beta(F_{n-1})$. Starting from $f_0(x) = x$, recursively define

$$f_n(x; x_1, \ldots, x_n) = f_{n-1}(x; x_1, \ldots, x_{n-1}) + \beta(f_n(x; x_1, \ldots, x_{n-1}))x_n.$$

Then $F_n = f_n(F_0; X_1, \ldots, X_n)$. Now define

$$g_n(x; x_1, \ldots, x_n) = \max_{0 \le k \le n} f_k(x; x_1, \ldots, x_k).$$

If $F_0 = x$, then $T_n(x) = [g_n(x; X_1, \ldots, X_n) = 1]$ is the event that bold play will by time n successfully increase the gambler's fortune to 1. From the recursive definition it follows by induction on n that for $n \ge 1$, $f_n(x; x_1, \ldots, x_n) = f_{n-1}(x + \beta(x)x_1; x_2, \ldots, x_n)$ and hence that $g_n(x; x_1, \ldots, x_n) = \max\{x, g_{n-1}(x + \beta(x)x_1; x_2, \ldots, x_n)\}$. Since $x = 1$ implies $g_{n-1}(x + \beta(x)x_1; x_2, \ldots, x_n) \ge x + \beta(x)x_1 = 1$, $T_n(x) = [g_{n-1}(x + \beta(x)X_1; X_2, \ldots, X_n) = 1]$, and since the X_i are independent and identically distributed, $P(T_n(x)) = P([X_1 = +1] \cap T_n(x)) + P([X_1 = -1] \cap T_n(x)) = pP[g_{n-1}(x + \beta(x); X_2, \ldots, X_n) = 1] + qP[g_{n-1}(x - \beta(x); X_2, \ldots, X_n) = pP(T_{n-1}(x + \beta(x))) + qP(T_{n-1}(x - \beta(x)))$. Letting $n \to \infty$ now gives $Q(x) = pQ(x + \beta(x)) + qQ(x - \beta(x))$, which reduces to (7.30) because $Q(0) = 0$ and $Q(1) = 1$.

Suppose that $y = f_{n-1}(x; x_1, \ldots, x_{n-1})$ is nondecreasing in x. If $x_n = +1$, then $f_n(x; x_1, \ldots, x_n)$ is $2y$ if $0 \le y \le \frac{1}{2}$ and 1 if $\frac{1}{2} \le y \le 1$; if $x_n = -1$, then $f_n(x; x_1, \ldots, x_n)$ is 0 if $0 \le y \le \frac{1}{2}$ and $2y - 1$ if $\frac{1}{2} \le y \le 1$. In any case, $f_n(x; x_1, \ldots, x_n)$ is also nondecreasing in x, and by induction this is true for every n. It follows that the same is true of $g_n(x; x_1, \ldots, x_n)$, of $P(T_n(x))$, and of $Q(x)$. Thus $Q(x)$ is nondecreasing.

Since $Q(1) = 1$, (7.30) implies that $Q\left(\frac{1}{2}\right) = pQ(1) = p$, $Q\left(\frac{1}{4}\right) = pQ\left(\frac{1}{2}\right) = p^2$, $Q\left(\frac{3}{4}\right) = p + qQ\left(\frac{1}{2}\right) = p + pq$. More generally, if $p_0 = p$ and $p_1 = q$, then

$$Q\left(\frac{k}{2^n}\right) = \sum \left[p_{u_1} \cdots p_{u_n} \colon \sum_{i=1}^{n} \frac{u_i}{2^i} < \frac{k}{2^n} \right], \quad 0 < k \le 2^n, \quad n \ge 1, \quad (7.31)$$

the sum extending over n-tuples $(u_1, \ldots, u_n)$ of 0's and 1's satisfying the condition indicated. Indeed, it is easy to see that (7.31) is the same thing as

$$Q(.u_1 \ldots u_n + 2^{-n}) - Q(.u_1 \ldots u_n) = p_{u_1} p_{u_2} \cdots p_{u_n} \quad (7.32)$$

for each dyadic rational $.u_1 \ldots u_n$ of rank n. If $.u_1 \ldots u_n + 2^{-n} \le \frac{1}{2}$, then $u_1 = 0$ and by (7.30) the difference in (7.32) is $p_0[Q(.u_2 \ldots u_n + 2^{-n+1}) - Q(.u_2 \ldots u_n)]$. But (7.32) follows inductively from this and a similar relation for the case $.u_1 \ldots u_n \ge \frac{1}{2}$.

Therefore $Q(k2^{-n}) - Q((k-1)2^{-n})$ is bounded by $\max\{p^n, q^n\}$, and so by monotonicity Q is continuous. Since (7.32) is positive, it follows that Q is strictly increasing over [0, 1].

Thus Q is continuous and increasing and satisfies (7.30). The inequality (7.28) is still to be proved. It is equivalent to the assertion that

$$\Delta(r,s) = Q(a) - pQ(s) - qQ(r) \geq 0$$

if $0 \leq r \leq s \leq 1$, where a stands for the average: $a = \frac{1}{2}(r+s)$. Since Q is continuous, it suffices to prove the inequality for r and s of the form $k/2^n$, and this will be done by induction on n. Checking all cases disposes of $n = 0$. Assume that the inequality holds for a particular n, and that r and s have the form $k/2^{n+1}$. There are four cases to consider.

Case 1. $s \leq \frac{1}{2}$. By the first part of (7.30), $\Delta(r,s) = p\Delta(2r, 2s)$. Since $2r$ and $2s$ have the form $k/2^n$, the induction hypothesis implies that $\Delta(2r, 2s) \geq 0$.

Case 2. $\frac{1}{2} \leq r$. By the second part of (7.30),

$$\Delta(r,s) = q\Delta(2r-1, 2s-1) \geq 0.$$

Case 3. $r \leq a \leq \frac{1}{2} \leq s$. By (7.30),

$$\Delta(r,s) = pQ(2a) - p[p + qQ(2s-1)] - q[pQ(2r)].$$

From $\frac{1}{2} \leq s \leq r+s = 2a \leq 1$, follows $Q(2a) = p + qQ(4a-1)$; and from $0 \leq 2a - \frac{1}{2} \leq \frac{1}{2}$, follows $Q\left(2a - \frac{1}{2}\right) = pQ(4a-1)$. Therefore, $pQ(2a) = p^2 + qQ\left(2a - \frac{1}{2}\right)$, and it follows that

$$\Delta(r,s) = q\left[Q\left(2a - \tfrac{1}{2}\right) - pQ(2s-1) - pQ(2r)\right].$$

Since $p \leq q$, the right side does not increase if either of the two p's is changed to q. Hence

$$\Delta(r,s) \geq q \max[\Delta(2r, 2s-1), \Delta(2s-1, 2r)].$$

The induction hypothesis applies to $2r \leq 2s-1$ or to $2s-1 \leq 2r$, as the case may be, so one of the two Δ's on the right is nonnegative.

Case 4. $r \leq \frac{1}{2} \leq a \leq s$. By (7.30),

$$\Delta(r,s) = pq + qQ(2a-1) - pqQ(2s-1) - pqQ(2r).$$

From $0 \le 2a-1 = r+s-1 \le \frac{1}{2}$, follows $Q(2a-1) = pQ(4a-2)$; and from $\frac{1}{2} \le 2a-\frac{1}{2} = r+s-\frac{1}{2} \le 1$, follows $Q\left(2a-\frac{1}{2}\right) = p+qQ(4a-2)$. Therefore, $qQ(2a-1) = pQ\left(2a-\frac{1}{2}\right) - p^2$, and it follows that

$$\Delta(r,s) = p\left[q-p+Q\left(2a-\frac{1}{2}\right) - qQ(2s-1) - qQ(2r)\right].$$

If $2s-1 \le 2r$, the right side here is

$$p[(q-p)(1-Q(2r)) + \Delta(2s-1, 2r)] \ge 0.$$

If $2r \le 2s-1$, the right side is

$$p[(q-p)(1-Q(2s-1)) + \Delta(2r, 2s-1)] \ge 0.$$

This completes the proof of (7.28) and hence of Theorem 7.3. ■

The equation (7.31) has an interesting interpretation. Let $Z_1, Z_2, \ldots$ be independent random variables satisfying $P[Z_n = 0] = p_0 = p$ and $P[Z_n = 1] = p_1 = q$. From $P[Z_n = 1 \quad \text{i.o.}] = 1$ and $\sum_{i>n} Z_i 2^{-i} \le 2^{-n}$ it follows that $P\left[\sum_{i=1}^{\infty} Z_i 2^{-i} \le k2^{-n}\right] \le P\left[\sum_{i=1}^{n} Z_i 2^{-i} < k2^{-n}\right] \le P\left[\sum_{i=1}^{\infty} Z_i 2^{-i} \le k2^{-n}\right]$. Since by (7.31) the middle term is $Q(k2^{-n})$,

$$Q(x) = P\left[\sum_{i=1}^{\infty} Z_i 2^{-i} \le x\right] \tag{7.33}$$

holds for dyadic rational x and hence by continuity holds for all x. In Section 31, Q will reappear as a continuous, strictly increasing function singular in the sense of Lebesgue. On p. 408 is a graph for the case $p_0 = .25$.

Note that $Q(x) \equiv x$ in the fair case $p = \frac{1}{2}$. In fact, for a bounded policy Theorem 7.2 implies that $E[F_\tau] = F_0$ in the fair case, and if the policy is to stop as soon as the fortune reaches 0 or 1, then the chance of successfully reaching 1 is $P[F_\tau = 1] = E[F_\tau] = F_0$. Thus in the fair case with initial fortune x, the chance of success is x for *every* policy that stops at the boundaries, and x is an upper bound even if stopping earlier is allowed.

EXAMPLE 7.8

The gambler of Example 7.1 has capital \$900 and goal \$1000. For a fair game $\left(p = \frac{1}{2}\right)$ his chance of success is .9 whether he bets unit stakes or adopts bold play. At red-and-black $\left(p = \frac{18}{38}\right)$, his chance of success with unit stakes is .00003; an approximate calculation based on (7.31) shows that under bold play his chance $Q(.9)$ of success increases to about .88, which compares well with the fair case.

EXAMPLE 7.9

In Example 7.2 the capital is \$100 and the goal \$20,000. At unit stakes the chance of successes is .005 for $p = \frac{1}{2}$ and 3×10^{-911} for $p = \frac{18}{38}$. Another approximate calculation shows that bold play at red-and-black gives the gambler probability about .003 of success, which again compares well with the fair case.

This example illustrates the point of Theorem 7.3. The gambler enters the casino knowing that he must by dawn convert his \$100 into \$20,000 or face certain death at the hands of criminals to whom he owes that amount. Only red-and-black is available to him. The question is not whether to gamble—he *must* gamble. The question is how to gamble so as to maximize the chance of survival, and bold play is the answer.

There are policies other than the bold one that achieve the maximum success probability $Q(x)$. Suppose that as long as the gambler's fortune x is less than $\frac{1}{2}$ he bets x for $x \le \frac{1}{4}$ and $\frac{1}{2} - x$ for $\frac{1}{4} \le x \le \frac{1}{2}$. This is, in effect, the bold-play strategy scaled down to the interval $\left[0, \frac{1}{2}\right]$, and so the chance he ever reaches $\frac{1}{2}$ is $Q(2x)$ for an initial fortune of x. Suppose further that if he does reach the goal of $\frac{1}{2}$, or if he starts with fortune at least $\frac{1}{2}$ in the first place, then he continues, but with ordinary bold play. For an initial fortune $x \ge \frac{1}{2}$, the overall chance of success is of course $Q(x)$, and for an initial fortune $x < \frac{1}{2}$, it is $Q(2x)Q\left(\frac{1}{2}\right) = pQ(2x) = Q(x)$. The success probability is indeed $Q(x)$ as for bold play, although the policy is different. With this example in mind, one can generate a whole series of distinct optimal policies.

Timid Play†

The optimality of bold play seems reasonable when one considers the effect of its opposite, timid play. Let the ϵ-*timid* policy be to bet $W_n = \min\{\epsilon, F_{n-1}, 1 - F_{n-1}\}$ and stop when F_n reaches 0 or 1. Suppose that $p < q$, fix an initial fortune $x = F_0$ with $0 \le x < 1$, and consider what happens as $\epsilon \to 0$. By the strong law of large numbers, $\lim_n n^{-1}S_n = E[X_1] = p - q < 0$. There is therefore probability 1 that $\sup_k S_k < \infty$ and $\lim_n S_n = -\infty$. Given $\eta > 0$, choose ϵ so that $P[\sup_k (x + \epsilon S_k) < 1] > 1 - \eta$. Since $P(\cup_{n=1}^{\infty}[x + \epsilon S_n < 0]) = 1$, with probability at least $1 - \eta$ there exists an n such that $x + \epsilon S_n < 0$ and $\max_{k<n}(x + \epsilon S_k) < 1$. But under the ϵ-timid policy the gambler is in this circumstance ruined. If $Q_\epsilon(x)$ is the probability of success under the ϵ-timid policy, then $\lim_{\epsilon \to 0} Q_\epsilon(x) = 0$ for $0 \le x < 1$. The law of large numbers carries the timid player to his ruin.‡

†This topic may be omitted.

‡For each ϵ, however, there exist optimal policies under which the bet never exceeds ϵ; see DUBINS & SAVAGE.

PROBLEMS

7.1. A gambler with initial capital a plays until his fortune increases b units or he is ruined. Suppose that $\rho > 1$. The chance of success is multiplied by $1 + \theta$ if his initial capital is infinite instead of a. Show that $0 < \theta < (\rho^a - 1)^{-1} < (a(\rho - 1))^{-1}$; relate to Example 7.3.

7.2. As shown on p. 94, there is probability 1 that the gambler either achieves his goal of c or is ruined. For $p \neq q$, deduce this directly from the strong law of large numbers. Deduce it (for all p) via the Borel–Cantelli lemma from the fact that if play never terminates, there can never occur c successive $+1$'s.

7.3. 6.12↑ If V_n is the set of n-long sequences of ± 1's, the function b_n in (7.9) maps V_{n-1} into $\{0, 1\}$. A selection system is a sequence of such maps. Although there are uncountably many selection systems, how many have an *effective* description in the sense of an algorithm or finite set of instructions by means of which a deputy (perhaps a machine) could operate the system for the gambler? An analysis of the question is a matter for mathematical logic, but one can see that there can be only countably many algorithms or finite sets of rules expressed in finite alphabets.

Let $Y_1^{(\sigma)}, Y_2^{(\sigma)}, \ldots$ be the random variables of Theorem 7.1 for a particular system σ, and let C_σ be the ω-set where every k-tuple of ± 1's (k arbitrary) occurs in $Y_1^{(\sigma)}(\omega), Y_2^{(\sigma)}(\omega), \ldots$ with the right asymptotic relative frequency (in the sense of Problem 6.12). Let C be the intersection of C_σ over all effective selection systems σ. Show that C lies in $\mathscr{F}$ (the σ-field in the probability space $(\Omega, \mathscr{F}, P)$ on which the X_n are defined) and that $P(C) = 1$. A sequence $(X_1(\omega), X_2(\omega), \ldots)$ for ω in C is called a *collective*: a subsequence chosen by any of the effective rules σ contains all k-tuples in the correct proportions.

7.4. Let D_n be 1 or 0 according as $X_{2n-1} \neq X_{2n}$ or not, and let M_k be the time of the kth 1—the smallest n such that $\sum_{i=1}^{n} D_i = k$. Let $Z_k = X_{2M_k}$. In other words, look at successive nonoverlapping pairs (X_{2n-1}, X_{2n}), discard accordant $(X_{2n-1} = X_{2n})$ pairs, and keep the second element of discordant $(X_{2n-1} \neq X_{2n})$ pairs. Show that this process simulates a fair coin: $Z_1, Z_2, \ldots$ are independent and identically distributed and $P[Z_k = +1] = P[Z_k = -1] = \frac{1}{2}$, whatever p may be. Follow the proof of Theorem 7.1.

7.5. Suppose that a gambler with initial fortune 1 stakes a proportion $\theta (0 < \theta < 1)$ of his current fortune: $F_0 = 1$ and $W_n = \theta F_{n-1}$. Show that $F_n =$

$\Pi_{k=1}^{n}(1+\theta X_k)$ and hence that

$$\log F_n = \frac{n}{2}\left[\frac{S_n}{n}\log\frac{1+\theta}{1-\theta} + \log(1-\theta^2)\right].$$

Show that $F_n \to 0$ with probability 1 in the subfair case.

7.6. In "doubling," $W_1 = 1, W_n = 2W_{n-1}$, and the rule is to stop after the first win. For any positive p, play is certain to terminate. Here $F_\tau = F_0 + 1$, but of course infinite capital is required. If $F_0 = 2^k - 1$ and W_n cannot exceed F_{n-1}, the probability of $F_\tau = F_0 + 1$ in the fair case is $1 - 2^{-k}$. Prove this via Theorem 7.2 and also directly.

7.7. In "progress and pinch," the wager, initially some integer, is increased by 1 after a loss and decreased by 1 after a win, the stopping rule being to quit if the next bet is 0. Show that play is certain to terminate if and only if $p \geq \frac{1}{2}$. Show that $F_\tau = F_0 + \frac{1}{2}W_1^2 + \frac{1}{2}(\tau - 1)$. Infinite capital is required.

7.8. Here is a common martingale. Just before the nth spin of the wheel, the gambler has before him a pattern $x_1, \ldots x_k$ of positive numbers (k varies with n). He bets $x_1 + x_k$, or x_1 in case $k = 1$. If he loses, at the next stage he uses the pattern $x_1, \ldots, x_k, x_1 + x_k$ (x_1, x_1 in case $k = 1$). If he wins, at the next stage he uses the pattern $x_2, \ldots, x_{k-1}$, unless k is 1 or 2, in which case he quits. Show that play is certain to terminate if $p > \frac{1}{3}$ and that the ultimate gain is the sum of the numbers in the initial pattern. Infinite capital is again required.

7.9. Suppose that $W_k = 1$, so that $F_k = F_0 + S_k$. Suppose that $p \geq q$ and τ is a stopping time such that $1 \leq \tau \leq n$ with probability 1. Show that $E[F_\tau] \leq E[F_n]$, with equality in case $p = q$. Interpret this result in terms of a stock option that must be exercised by time n, where $F_0 + S_k$ represents the price of the stock at time k.

7.10. For a given policy, let A_n^* be the fortune of the gambler's adversary at time n. Consider these conditions on the policy: (i) $W_n^* \leq F_{n-1}^*$; (ii) $W_n^* \leq A_{n-1}^*$; (iii) $F_n^* + A_n^*$ is constant. Interpret each condition, and show that together they imply that the policy is bounded in the sense of (7.24).

7.11. Show that F_τ has infinite range if $F_0 = 1, W_n = 2^{-n}$, and τ is the smallest n for which $X_n = +1$.

7.12. Let u be a real function on [0, 1], $u(x)$ representing the *utility* of the fortune x. Consider policies bounded by 1; see (7.24). Let $Q_\pi(F_0) = E[u(F_\tau)]$: this represents the expected utility under the policy π of an initial fortune F_0. Suppose of a policy π_0 that

$$u(x) \leq Q_{\pi_0}(x), \quad 0 \leq x \leq 1, \tag{7.34}$$

and that

$$Q_{\pi_0}(x) \geq pQ_{\pi_0}(x+t) + qQ_{\pi_0}(x-t), \qquad (7.35)$$
$$0 \leq x-t \leq x \leq x+t \leq 1.$$

Show that $Q_\pi(x) \leq Q_{\pi_0}(x)$ for all x and all policies π. Such a π_0 is optimal.

Theorem 7.3 is the special case of this result for $p \leq \frac{1}{2}$, bold play in the role of π_0, and $u(x) = 1$ or $u(x) = 0$ according as $x = 1$ or $x < 1$.

The condition (7.34) says that gambling with policy π_0 is at least as good as not gambling at all; (7.35) says that, although the prospects even under π_0 become on the average less sanguine as time passes, it is better to use π_0 now than to use some other policy for one step and then change to π_0.

7.13. The functional equation (7.30) and the assumption that Q is bounded suffice to determine Q completely. First, $Q(0)$ and $Q(1)$ must be 0 and 1, respectively, and so (7.31) holds. Let $T_0x = \frac{1}{2}x$ and $T_1x = \frac{1}{2}x + \frac{1}{2}$; let $f_0x = px$ and $f_1x = p + qx$. Then $Q(T_{u_1} \cdots T_{u_n}x) = f_{u_1} \cdots f_{u_n}Q(x)$. If the binary expansions of x and y both begin with the digits $u_1, \ldots u_n$, they have the form $x = T_{u_1} \cdots T_{u_n}x'$ and $y = T_{u_1} \cdots T_{u_n}y'$. If K bounds Q and if $m = \max\{p, q\}$, it follows that $|Q(x) - Q(y)| \leq Km^n$. Therefore, Q is continuous and satisfies (7.31) and (7.33)

SECTION 8 MARKOV CHAINS

As Markov chains illustrate in a clear and striking way the connection between probability and measure, their basic properties are developed here in a measure-theoretic setting.

Definitions

Let S be a finite or countable set. Suppose that to each pair i and j in S there is assigned a nonnegative number p_{ij} and that these numbers satisfy the constraint

$$\sum_{j \in S} p_{ij} = 1, \quad i \in S. \qquad (8.1)$$

Let $X_0, X_1, X_2, \ldots$ be a sequence of random variables whose ranges are contained in S. The sequence is a *Markov chain* or *Markov process* if

$$P[X_{n+1} = j \,[X_0 = t_0, \ldots, X_n = i_n] \qquad (8.2)$$
$$= P[X_{n+1} = j \,|X_n = i_n] = p_{i_n j}$$

for every n and every sequence $i_0, \ldots, i_n$ in S for which $P[X_0 = i_0, \ldots, X_n = i_n] > 0$. The set S is the *state space* or *phase space* of the process, and the p_{ij} are the *transition probabilities*. Part of the defining condition (8.2) is that the transition probability

$$P[X_{n-1} = j \,|\, X_n = j] = p_{ij} \tag{8.3}$$

does not vary with n.[†]

The elements of S are thought of as the possible states of a *system*, X_n representing the state at *time* n. The sequence or process $X_0, X_1, X_2, \ldots$ then represents the history of the system, which evolves in accordance with the probability law (8.2). The conditional distribution of the *next* state X_{n+1} given the *present* state X_n must not further depend on the *past* $X_0, \ldots X_{n-1}$. This is what (8.2) requires, and it leads to a copious theory.

The *initial probabilities* are

$$\alpha_i = P[X_0 = i]. \tag{8.4}$$

The α_i are nonnegative and add to 1, but the definition of Markov chain places no further restrictions on them.

The following examples illustrate some of the possibilities. In each one, the state space S and the transition probabilities p_{ij} are described, but the underlying probability space $(\Omega, \mathscr{F}, P)$ and the X_n are left unspecified for now: see Theorem 8.1.[‡]

EXAMPLE 8.1

The Bernoulli–Laplace model of diffusion. Imagine r black balls and r white balls distributed between two boxes, with the constraint that each box contains r balls. The state of the system is specified by the number of white balls in the first box, so that the state space is $S = \{0, 1, \ldots, r\}$. The transition mechanism is this: at each stage one ball is chosen at random from each box and the two are interchanged. If the present state is i, the chance of a transition to $i-1$ is the chance i/r of drawing one of the i white balls from the first box times the chance i/r of drawing one of the i black balls from the second box. Together with similar arguments for the other possibilities, this shows that the transition

[†]Sometimes in the definition of the Markov chain $P[X_{n+1} = j \,|\, X_n = i]$ is allowed to depend on n. A chain satisfying (8.3) is then said to have *stationary transition probabilities*, a phrase that will be omitted here because (8.3) will always be assumed.

[‡]For an excellent collection of examples from physics and biology, see FELLER, Volume 1. Chapter XV.

probabilities are

$$p_{i,i-1} = \left(\frac{i}{r}\right)^2, \qquad p_{i,i+1} = \left(\frac{r-i}{r}\right)^2, \qquad p_{ii} = 2\frac{i(r-i)}{r^2},$$

the others being 0. This is the probablistic analogue of the model for the flow of two liquids between two containers.

The p_{ij} form the *transition matrix* $P = [p_{ij}]$ of the process. A *stochastic matrix* is one whose entries are nonnegative and satisfy (8.1); the transition matrix of course has this property.

EXAMPLE 8.2

Random walk with absorbing barriers. Suppose that $S = \{0, 1, \ldots, r\}$ and

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & q & 0 & p & \ldots & 0 & 0 & 0 & 0 \\ \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. \\ 0 & 0 & 0 & 0 & \ldots & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 \end{bmatrix}$$

That is, $p_{i,i+1} = p$ and $p_{i,i-1} = q = 1 - p$ for $0 < i < r$ and $p_{00} = p_{rr} = 1$. The chain represents a particle in *random walk*. The particle moves one unit to the right or left, the respective probabilities being p and q, except that each of 0 and r is an *absorbing* state—once the particle enters, it cannot leave. The state can also be viewed as a gambler's fortune; absorption in 0 represents ruin for the gambler, absorption in r ruin for his adversary (see Section 7). The gambler's initial fortune is usually regarded as nonrandom, so that (see (8.4)) $\alpha_i = 1$ for some i.

EXAMPLE 8.3

Unrestricted random walk. Let S consist of all the integers $i = 0, \pm 1, \pm 2, \ldots$, and take $p_{i,i+1} = p$ and $p_{i,i-1} = q = 1 - p$. This chain represents a random walk without barriers, the particle being free to move anywhere on the integer lattice. The walk is *symmetric* if $p = q$.

The state space may, as in the preceding example, be countably infinite. If so, the Markov chain consists of functions X_n on a probability space $(\Omega, \mathscr{F}, P)$,

but these will have infinite range and hence will not be random variables in the sense of the preceding sections. This will cause no difficulty, however, because expected values of the X_n will not be considered. All that is required is that for each $i \in S$ the set $[\omega\colon X_n(\omega) = i]$ lie in $\mathscr{F}$ and hence have a probability.

EXAMPLE 8.4

Symmetric random walk in space. Let S consist of the integer lattice points in k-dimensional Euclidean space R^k; $x = (x_1, \ldots, x_k)$ lies in S if the coordinates are all integers. Now x has $2k$ neighbors, points of the form $y = (x_1, \ldots, x_i \pm 1, \ldots, x_k)$; for each such y let $p_{xy} = (2k)^{-1}$. The chain represents a particle moving randomly in space; for $k = 1$ it reduces to Example 8.3 with $p = q = \frac{1}{2}$. The cases $k \leq 2$ and $k \geq 3$ exhibit an interesting difference. If $k \leq 2$, the particle is certain to return to its initial position, but this is not so if $k \geq 3$; see Example 8.6.

Since the state space in this example is not a subset of the line, the $X_0, X_1, \ldots$ do not assume real values. This is immaterial because expected values of the X_n play no role. All that is necessary is that X_n be a mapping from Ω into S (finite or countable) such that $[\omega\colon X_n(\omega) = i] \in \mathscr{F}$ for $i \in S$. There will be expected values $E[f(X_n)]$ for real functions f on S with finite range, but then $f(X_n(\omega))$ is a simple random variable as defined before.

EXAMPLE 8.5

A selection problem. A princess must chose from among r suitors. She is definite in her preferences and if presented with all r at once could choose her favorite and could even rank the whole group. They are ushered into her presence one by one in random order, however, and she must at each stage either stop and accept the suitor or else reject him and proceed in the hope that a better one will come along. What strategy will maximize her chance of stopping with the best suitor of all?

Shorn of some details, the analysis is this. Let $S_1, S_2, \ldots, S_r$ be the suitors in order of presentation; this sequence is a random permutation of the set of suitors. Let $X_1 = 1$ and let $X_2, X_3, \ldots$ be the successive positions of suitors who dominate (are preferable to) all their predecessors. Thus $X_2 = 4$ and $X_3 = 6$ means that S_1 dominates S_2 and S_3 but S_4 dominates S_1, S_2, S_3, and that S_4 dominates S_5 but S_6 dominates $S_1, \ldots, S_5$. There can be at most r of these dominant suitors; if there are exactly m, $X_{m+1} = X_{m+2} = \cdots = r + 1$ by convention.

As the suitors arrive in random order, the chance that S_i ranks highest among $S_1, \ldots, S_i$ is $(i-1)!/i! = 1/i$. The chance that S_j ranks highest among

$S_1, \ldots, S_j$ and S_i ranks next is $(j-2)!/j! = 1/j(j-1)$. This leads to a chain with transition probabilities[†]

$$P[X_{n+1} = j | X_n = i] = \frac{i}{j(j-1)}, \quad 1 \le i < j \le r. \tag{8.5}$$

If $X_n = i$, then $X_{n+1} = r+1$ means that S_i dominates $S_{i+1}, \ldots, S_r$ as well as $S_1, \ldots, S_i$, and the conditional probability of this is

$$P[X_{n+1} = r+1 | X_n = i] = \frac{1}{r}, \quad 1 \le i \le r. \tag{8.6}$$

As downward transitions are impossible and $r+1$ is absorbing, this specifies a transition matrix for $S = \{1, 2, \ldots, r+1\}$.

It is quite clear that in maximizing her chance of selecting the best suitor of all, the princess should reject those who do not dominate their predecessors. Her strategy therefore will be to stop with the suitor in position X_r, where τ is a random variable representing her strategy. Since her decision to stop must depend only on the suitors she has seen thus far, the event $[\tau = n]$ must lie in $\sigma(X, \ldots, X_n)$. If $X_\tau = i$, then by (8.6) the conditional probability of success is $f(i) = i/r$. The probability of success is therefore $E[f(X_\tau)]$, and the problem is to choose the strategy τ so as to maximize it. For the solution, see Example 8.17.[‡]

Higher-Order Transitions

The properties of the Markov chain are entirely determined by the transition and initial probabilities. The chain rule (4.2) for conditional probabilities gives

$$\begin{aligned} P[X_0 = i_0, X_1 = i_1, X_2 = i_2] &= P[X_0 = i_0] P[X_1 = i_1 | X_0 = i_0] P[X_2 = i_2 | X_0 = i_0, X_1 = i_1] \\ &= \alpha_{i_0} p_{i_0 t_1} p_{i_1 i_2}. \end{aligned}$$

Similarly,

$$P[X_t = i_t, 0 \le t \le m] = \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{m-1} i_m} \tag{8.7}$$

for any sequence $i_0, i_1, \ldots, i_m$ of states.

Further,

$$P[X_{m+t} = J_t, 1 \le t \le n | X_s i_s, 0 \le s \le m] = p_{i_m j_1} p_{j_1 j_2} \cdots p_{j_{n-1} j_n}, \tag{8.8}$$

[†]The details can be found in DYNKIN & YUSHKEVICH. Chapter III.

[‡]With the princess replaced by an executive and the suitors by applicants for an office job, this is known as the *secretary problem.*

as follows by expressing the conditional probability as a ratio and applying (8.7) to numerator and denominator. Adding out the intermediate states now gives the formula

$$p_{ij}^{(n)} = P[X_{m+n} = j \,|X_m = i] \tag{8.9}$$
$$= \sum_{k_1 \ldots k_{n-1}} p_{ik_1} p_{k_1 k_2} \cdots p_{k_{n-1} j}$$

(the k_1 range over S) for the *nth-order transition probabilities*.

Notice that $p_{ij}^{(n)}$ is the entry in position (i, j) of P^n, the nth power of the transition matrix P. If S is infinite, P is a matrix with infinitely many rows and columns; as the terms in (8.9) are nonnegative, there are no convergence problems. It is natural to put

$$p_{ij}^{(0)} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Then P^0 is the identity I, as it should be. From (8.1) and (8.9) follow

$$p_{ij}^{(m+n)} = \sum_v p_{iv}^{(m)} p_{vj}^{(n)}, \qquad \sum_j p_{ij}^{(n)} = 1. \tag{8.10}$$

An Existence Theorem

THEOREM 8.1

Suppose that $P = [p_{ij}]$ is a stochastic matrix and that α_i are nonnegative numbers satisfying $\sum_{i \in S} \alpha_i = 1$. There exists on some $(\Omega, \mathscr{F}, P)$ a Markov chain $X_0, X_1, X_2, \ldots$ with initial probabilities α_i and transition probabilities p_{ij}.

Proof. Reconsider the proof of Theorem 5.3. There the space $(\Omega, \mathscr{F}, P)$ was the unit interval, and the central part of the argument was the construction of the decompositions (5.13). Suppose for the moment that $S = \{1, 2, \ldots\}$. First construct a partition $I_1^{(0)}, I_2^{(0)}, \ldots$ of (0, 1] into countably many† subintervals of lengths (P is again Lebesgue measure) $P(I_i^{(0)}) = \alpha_i$. Next decompose each $I_i^{(0)}$ into subintervals $I_{ij}^{(1)}$ of lengths $P(I_{ij}^{(1)}) = \alpha_i p_{ij}$. Continuing inductively gives a sequence of partitions $\{I_{i_0 \ldots i_n}^{(n)} : i_0, \ldots, i_n = 1, 2, \ldots\}$ such that each refines the preceding and $P(I_{i_0 \ldots i_n}^{(n)}) = \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}$.

Put $X_n(\omega) = i$ if $\omega \in \bigcup_{i_0 \ldots i_{n-1}} I_{i_0 \ldots i_{n-1} i}^{(n)}$. It follows just as in the proof of Theorem 5.3 that the set $[X_0 = i_0, \ldots, X_n = i_n]$ coincides with the interval $I_{i_0}^{(n)} \ldots i_n$. Thus $P[X_0 = i_0, \ldots, X_n = i_n] = \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}$. From this it follows immediately that (8.4) holds and that the first and third members of (8.2)

†If $\delta_1 + \delta_2 + \cdots = b - a$ and $\delta_i \geq 0$, then $I_i = (b - \sum_{j \leq i} \delta_j, b - \sum_{j < i} \delta_j]$, $i = 1, 2, \ldots$, decompose $(a, b]$ into intervals of lengths δ_i.

are the same. As for the middle member, it is $P[X_n = i_n, X_{n+1} = j]/P[X_n = i_n]$; the numerator is $\sum \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n} p_{i_n j}$ the sum extending over all $i_0, \ldots, i_{n-1}$, and the denominator is the same thing without the factor $p_{i_n j}$, which means that the ratio is $p_{i_n j}$, as required.

That completes the construction for the case $S = \{1, 2, \ldots\}$. For the general countably infinite S, let g be a one-to-one mapping of $\{1, 2, \ldots\}$ onto S, and replace the X_n as already constructed by $g(X_n)$; the assumption $S = \{1, 2, \ldots\}$ was merely a notational convenience. The same argument obviously works if S is finite.[†] ■

Although strictly speaking the Markov chain *is* the sequence $X_0, X_1, \ldots$, one often speaks as though the chain were the matrix P together with the initial probabilities α_i or even P with some unspecified set of α_i. Theorem 8.1 justifies this attitude: For given P and α_i the corresponding X_n do exist, and the apparatus of probability theory—the Borel–Cantelli lemmas and so on—is available for the study of P and of systems evolving in accordance with the Markov rule.

From now on fix a chain $X_0, X_1, \ldots$ satisfying $\alpha_i > 0$ for all i. Denote by P_i probabilities conditional on $[X_0 = i]$: $P_i(A) = P[A|X_0 = i]$. Thus

$$P_i[X_t = i_t, 1 \le t \le n] = p_{i i_1} p_{i_1, i_2} \cdots p_{i_{n-1} i_n} \tag{8.11}$$

by (8.8). The interest centers on these conditional probabilities, and the actual initial probabilities α_i are now largely irrelevant.

From (8.11) follows

$$\begin{aligned} &P_i[X_1 = i_1, \ldots, X_m = i_m, X_{m+1} = j_1, \ldots, X_{m+n} = j_n] \\ &\quad = P_i[X_1 = i_1, \ldots, X_m = i_m] P_{i_m}[X_1 = j_1 \ldots, X_n = j_n]. \end{aligned} \tag{8.12}$$

Suppose that I is a set (finite or infinite) of m-long sequences of states, J is a set of n-long sequences of states, and every sequence in I ends in j. Adding both sides of (8.12) for $(i_1, \ldots, i_m)$ ranging over I and $(j_1, \ldots, j_n)$ ranging over J gives

$$\begin{aligned} &P_i[(X_1, \ldots, X_m) \in I, (X_{m+1}, \ldots, X_{m+n}) \in J] \\ &\quad = P_i[(X_1, \ldots, X_m) \in I] P_j[(X_1, \ldots, X_n) \in J]. \end{aligned} \tag{8.13}$$

For this to hold it is essential that each sequence in I end in j. The formulas (8.12) and (8.13) are of central importance.

[†]For a different approach in the finite case, see Problem 8.1.

Transience and Persistence

Let

$$f_{ij}^{(n)} = P_i[X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j] \tag{8.14}$$

be the probability of a first visit to j at time n for a system that starts in i, and let

$$f_{ij} = P_i(\bigcup_{n=1}^{\infty}[X_n = j]) = \sum_{n=1}^{\infty} f_{ij}^{(n)} \tag{8.15}$$

be the probability of an eventual visit. A state i is *persistent* if a system starting at i is certain sometime to return to i: $f_{ii} = 1$. The state is *transient* in the opposite case: $f_{ii} < 1$.

Suppose that $n_1, \ldots, n_k$ are integers satisfying $1 \leq n_1 < \cdots < n_k$ and consider the event that the system visits j at times $n_1 \ldots n_k$ but not in between; this event is determined by the conditions

$$\begin{array}{lll} X_1 \neq j, \ldots, & X_{n_1-1} \neq j, & X_{n_1} = j, \\ X_{n_1+1} \neq j, \ldots, & X_{n_2} - 1 \neq j, & X_{n_2} = j, \\ & \vdots & \\ X_{n_k-1} + 1 \neq j, \ldots, & X_{n_k-1} \neq j, & X_{n_k} = j. \end{array}$$

Repeated application of (8.13) shows that under P_i the probability of this event is $f_{ij}^{(n_1)} f_{ij}^{(n_2-n_1)} \cdots f_{ij}^{(n_k-n_{k-1})}$. Add this over the k-tuples $n_1, \ldots, n_k$: the P_i-probability that $X_n = j$ for at least k different values of n is $f_{ij} f_{ij}^{k-1}$. Letting $k \to \infty$ therefore gives

$$P_i[X_n = j \text{ i.o.}] = \begin{cases} 0 & \text{if } f_{ij} < 1, \\ f_{ij} & \text{if } f_{ij} = 1. \end{cases} \tag{8.16}$$

Recall that *i.o.* means *infinitely often*. Taking $i = j$ gives

$$P_i[X_n = i \text{ i.o.}] = \begin{cases} 0 & \text{if } f_{ii} < 1, \\ f_{ii} & \text{if } f_{ii} = 1. \end{cases} \tag{8.17}$$

Thus $P_i[X_n = i \text{ i.o.}]$ is either 0 or 1; compare the zero–one law (Theorem 4.5), but note that the events $[X_n = i]$ here are not in general independent.[†]

THEOREM 8.2

(i) *Transience of i is equivalent to $P_i[X_n = i \text{ i.o.}] = 0$ and to $\sum_n p_{ii}^{(n)} < \infty$.*

(ii) *Persistence of i is equivalent to $P_i[X_n = i \text{ i.o.}] = 1$ and to $\sum_n p_{ii}^{(n)} = \infty$.*

[†]See Problem 8.35.

Proof. By the first Borel–Cantelli lemma, $\sum_n p_{ii}^{(n)} < \infty$ implies $P_i[X_n = i \text{ i.o.}] = 0$, which by (8.17) in turn implies $f_{ii} < 1$. The entire theorem will be proved if it is shown that $f_{ii} < 1$ implies $\sum_n p_{ii}^{(n)} < \infty$.

The proof uses a first-passage argument: By (8.13),

$$\begin{aligned} p_{ij}^{(n)} = P_i[X_n = j] &= \sum_{s=0}^{n-1} P_i[X_1 \neq j, \ldots, X_{n-s-1} \neq j, X_{n-s} = j, X_n = j] \\ &= \sum_{s=0}^{n-1} P_i[X_1 \neq j, \ldots, X_{n-s-1} \neq j, X_{n-s} = j] P_j[X_s = j] \\ &= \sum_{s=0}^{n-1} f_{ij}^{(n-s)} p_{jj}^{(s)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{t=1}^{n} p_{ii}^{(t)} &= \sum_{t=1}^{n} \sum_{s=0}^{t-1} f_{ii}^{(t-s)} p_{ii}^{(s)} \\ &= \sum_{s=0}^{n-1} p_{ii}^{(s)} \sum_{t=s+1}^{n} f_{ii}^{(t-s)} \leq \sum_{s=0}^{n} p_{ii}^{(s)} f_{ii}. \end{aligned}$$

Thus $(1 - f_{ii}) \sum_{i=1}^{n} p_{ii}^{(t)} \leq f_{ii}$, and if $f_{ii} < 1$, this puts a bound on the partial sums $\sum_{t=1}^{n} p_{ii}^{(t)}$. ■

EXAMPLE 8.6

Pólya's theorem. For the symmetric k-dimensional random walk (Example 8.4), all states are persistent if $k = 1$ or $k = 2$, and all states are transient if $k \geq 3$. To prove this, note first that the probability $p_{ii}^{(n)}$ of return in n steps is the same for all states i; denote this probability by $a_n^{(k)}$ to indicate the dependence on the dimension k. Clearly, $a_{2n+1}^{(k)} = 0$. Suppose that $k = 1$. Since return in $2n$ steps means n steps east and n steps west,

$$a_{2n}^{(1)} = \binom{2n}{n} \frac{1}{2^{2n}}.$$

By Stirling's formula, $a_{2n}^{(1)} \sim (\pi n)^{-1/2}$. Therefore, $\sum_n a_n^{(1)} = \infty$, and all states are persistent by Theorem 8.2.

In the plane, a return to the starting point in $2n$ steps means equal numbers of steps east and west as well as equal numbers north and south:

$$a_{2n}^{(2)} = \sum_{u=0}^{n} \frac{(2n)!}{u!u!(n-u)!(n-u)!} \frac{1}{4^{2n}}$$

$$= \frac{1}{4^{2n}} \binom{2n}{n} \sum_{u=0}^{n} \binom{n}{u} \binom{n}{n-u}.$$

It can be seen on combinatorial grounds that the last sum is $\binom{2n}{n}$, and so $a_{2n}^{(2)} = (a_{2n}^{(1)})^2 \sim (\pi n)^{-1}$. Again, $\sum_n a_n^{(2)} = \infty$ and every state is persistent.

For three dimensions,

$$a_{2n}^{(3)} = \sum \frac{(2n)!}{u!u!v!v!(n-u-v)!(n-u-v)!} \frac{1}{6^{2n}},$$

the sum extending over nonnegative u and v satisfying $u + v \le n$. This reduces to

$$a_{2n}^{(3)} = \sum_{i=0}^{n} \binom{2n}{2l} \left(\frac{1}{3}\right)^{2n-2l} \left(\frac{2}{3}\right)^{2l} a_{2n-2l}^{(1)} a_{2l}^{(2)}, \tag{8.18}$$

as can be checked by substitution. (To see the probabilistic meaning of this formula, condition on there being $2n - 2l$ steps parallel to the vertical axis and $2l$ steps parallel to the horizontal plane.) It will be shown that $a_{2n}^{(3)} = O(n^{-3/2})$, which will imply that $\sum_n a_n^{(3)} < \infty$. The terms in (8.18) for $l = 0$ and $l = n$ are each $O(n^{-3/2})$ and hence can be omitted. Now $a_u^{(1)} \le Ku^{-1/2}$ and $a_u^{(2)} \le Ku^{-1}$, as already seen, and so the sum in question is at most

$$K^2 \sum_{l=1}^{n-1} \binom{2n}{2l} \left(\frac{1}{3}\right)^{2n-2l} \left(\frac{2}{3}\right)^{2l} (2n-2l)^{-1/2} (2l)^{-1}.$$

Since $(2n-2l)^{-1/2} \le 2n^{1/2}(2n-2l)^{-1} \le 4n^{1/2}(2n-2l+1)^{-1}$ and $(2l)^{-1} \le 2(2l+1)^{-1}$, this is at most a constant times

$$n^{1/2} \frac{(2n)!}{(2n+2)!} \sum_{l=1}^{n-1} \binom{2n+2}{2l-1} \left(\frac{1}{3}\right)^{2n-2l+1} \left(\frac{2}{3}\right)^{2l+1} = O(n^{-3/2}).$$

Thus $\sum_n a_n^{(3)} < \infty$, and the states are transient. The same is true for $k = 4, 5, \ldots$, since an inductive extension of the argument shows that $a_n^{(k)} = O(n^{-k/2})$.

It is possible for a system starting in i to reach j ($f_{ij} > 0$) if and only if $p_{ij}^{(n)} > 0$ for some n. If this is true for all i and j, the Markov chain is *irreducible*.

THEOREM 8.3

If the Markov chain is irreducible, then one of the following two alternatives holds.

(i) *All states are transient,* $P_i(\bigcup_j [X_n = j \text{ i.o.}]) = 0$ *for all* i*, and* $\sum_n p_{ij}^{(n)} < \infty$ *for all* i *and* j*.*

(ii) *All states are persistent,* $P_i(\bigcap_j [X_n = j \text{ i.o.}]) = 1$ *for all* i*, and* $\sum_n p_{ij}^{(n)} = \infty$ *for all* i *and* j*.*

The irreducible chain itself can accordingly be called persistent or transient. In the persistent case the system visits every state infinitely often. In the transient case it visits each state only finitely often, hence visits each finite set only finitely often, and so may be said to go to infinity.

Proof. For each i and j there exist r and s such that $p_{ij}^{(r)} > 0$ and $p_{ji}^{(s)} > 0$. Now

$$p_{ii}^{(r+s+n)} \geq p_{ij}^{(r)} p_{ij}^{(n)} p_{ji}^{(s)}, \tag{8.19}$$

and from $p_{ij}^{(r)} p_{ji}^{(s)} > 0$ it follows that $\sum_n p_{ii}^{(n)} < \infty$ implies $\sum_n p_{jj}^{(n)} < \infty$: if one state is transient, they all are. In this case (8.16) gives $P_i[X_n = j \text{ i.o.}] = 0$ for all i and j, so that $P_i(\bigcup_j [X_n = j \text{ i.o.}]) = 0$ for all i. Since $\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{v=1}^{n} f_{ij}^{(v)} p_{jj}^{(n-v)} = \sum_{v=1}^{\infty} f_{ij}^{(v)} \sum_{m=0}^{\infty} p_{jj}^{(m)} \leq \sum_{m=0}^{\infty} p_{jj}^{(m)}$, it follows that if j is transient, then (Theorem 8.2) $\sum_n p_{ij}^{(n)}$ converges for every i.

The other possibility is that all states are persistent. In this case $P_j[X_n = j \text{ i.o.}] = 1$ by Theorem 8.2, and it follows by (8.13) that

$$\begin{aligned} p_{ji}^{(m)} &= P_j([X_m = i] \cap [X_n = j \text{ i.o.}]) \\ &\leq \sum_{n > m} P_j[X_m = i, X_{m+1} \neq j, \ldots, X_{n-1} \neq j, X_n = j] \\ &= \sum_{n > m} p_{ji}^{(m)} f_{ij}^{(n-m)} = p_{ji}^{(m)} f_{ij}. \end{aligned}$$

There is an m for which $p_{ji}^{(m)} > 0$, and therefore $f_{ij} = 1$. By (8.16), $P_i[X_n = j \text{ i.o.}] = f_{ij} = 1$. If $\sum_n p_{ij}^{(n)}$ were to converge for some i and j, it would follow by the first Borel–Cantelli lemma that $P_i[X_n = j \text{ i.o.}] = 0$. ■

EXAMPLE 8.7
Since $\sum_j p_{ij}^{(n)} = 1$, the first alternative in Theorem 8.3 is impossible if S is finite: a finite, irreducible Markov chain is persistent.

EXAMPLE 8.8
The chain in Pólya's theorem is certainly irreducible. If the dimension is 1 or 2, there is probability 1 that a particle in symmetric random walk visits every state infinitely often. If the dimension is 3 or more, the particle goes to infinity.

EXAMPLE 8.9

Consider the unrestricted random walk on the line (Example 8.3). According to the ruin calculation (7.8), $f_{01} = p/q$ for $p < q$. Since the chain is irreducible, all states are transient. By symmetry, of course, the chain is also transient if $p > q$, although in this case (7.8) gives $f_{01} = 1$. Thus $f_{ij} = 1 (i \neq j)$ is possible in the transient case.[†]

If $p = q = \frac{1}{2}$, the chain is persistent by Pólya's theorem. If n and $j-i$ have the same parity,

$$p_{ij}^{(n)} = \binom{n}{\frac{n+j-i}{2}} \frac{1}{2^n}, \quad |j - i| \leq n.$$

This is maximal if $j = i$ or $j = i \pm 1$, and by Stirling's formula the maximal value is of order $n^{-1/2}$. Therefore, $\lim_n p_{ij}^{(n)} = 0$, which always holds in the transient case but is thus possible in the persistent case as well (see Theorem 8.8).

Another Criterion for Persistence

Let $Q = [q_{ij}]$ be a matrix with rows and columns indexed by the elements of a finite or countable set U. Suppose it is *substochastic* in the sense that $q_{ij} \geq 0$ and $\sum_j q_{ij} \leq 1$. Let $Q^n = [q_{ij}^{(n)}]$ be the nth power, so that

$$q_{ij}^{(n+1)} = \sum_v q_{iv} q_{vj}^{(n)}, \quad q_{ij}^{(0)} = \delta_{ij}. \tag{8.20}$$

Consider the row sums

$$\sigma_i^{(n)} = \sum_j q_{ij}^{(n)}. \tag{8.21}$$

From (8.20) follows

$$\sigma_i^{(n+1)} = \sum_j q_{ij} \sigma_j^{(n)}. \tag{8.22}$$

Since Q is substochastic $\sigma_i^{(1)} \leq 1$, and hence $\sigma_i^{(n+1)} = \sum_j \sum_v q_{iv}^{(n)} q_{vj} = \sum_v q_{iv}^{(n)} \sigma_v^{(1)} \leq \sigma_i^{(n)}$. Therefore, the monotone limits

$$\sigma_i = \lim_n \sum_j q_{ij}^{(n)} \tag{8.23}$$

[†]But for each j there must be some $i \neq j$ for which $f_{ij} < 1$; see Problem 8.7.

exist. By (8.22) and the Weierstrass M-test [A28], $\sigma_i = \sum_j q_{ij}\sigma_j$. Thus the σ_i solve the system

$$\begin{cases} x_i = \sum\limits_{j\in U} q_{ij}x_j, & i \in U, \\ 0 \le x_i \le 1, & i \in U. \end{cases} \tag{8.24}$$

For an arbitrary solution, $x_i = \sum_j q_{ij}x_j \le \sum_j q_{ij} = \sigma_i^{(1)}$, and $x_i \le \sigma_i^{(n)}$ for all i implies $x_i \le \sum_j q_{ij}\sigma_j^{(n)} = \sigma_i^{(n+1)}$ by (8.22). Thus $x_i \le \sigma_i^{(n)}$ for all n by induction, and so $x_i \le \sigma_i$. Thus the σ_i give the *maximal* solution to (8.24):

Lemma 1. *For a substochastic matrix Q the limits* (8.23) *are the maximal solution of* (8.24)

Now suppose that U is a subset of the state space S. The p_{ij} for i and j in U give a substochastic matrix Q. The row sums (8.21) are $\sigma_i^{(n)} = \sum p_{ij_1}p_{j_1j_2}\cdots p_{j_{n-1}j_n}$, where the $j_1, \ldots, j_n$ range over U, and so $\sigma_i^{(n)} = P_i[X_t \in U, t \le n]$. Let $n \to \infty$:

$$\sigma_i = P_i[X_t \in U, t = 1, 2 \ldots], \quad i \in U. \tag{8.25}$$

In this case, σ_i is thus the probability that the system remains forever in U, given that it starts at i. The following theorem is now an immediate consequence of Lemma 1.

THEOREM 8.4
For $U \subset S$ the probabilities (8.25) *are the maximal solution of the system*

$$\begin{cases} x_i = \sum\limits_{j\in U} p_{ij}x_j, & i \in U, \\ 0 \le x_i \le 1, & i \in U, \end{cases} \tag{8.26}$$

The constraint $x_i \ge 0$ in (8.26) is in a sense redundant: Since $x_i \equiv 0$ is a solution, the maximal solution is automatically nonnegative (and similarly for (8.24)). And the maximal solution is $x_i \equiv 1$ if and only if $\sum_{j\in U} p_{ij} = 1$ for all i in U, which makes probabilistic sense.

EXAMPLE 8.10
For the random walk on the line consider the set $U = \{0, 1, 2, \ldots\}$. The system (8.26) is

$$x_i = px_{i+1} + qx_{i-1}, \quad i \ge 1,$$
$$x_0 = px_1.$$

It follows [A19] that $x_n = A + An$ if $p = q$ and $x_n = A - A(q/p)^{n+1}$ if $p \neq q$. The only bounded solution is $x_n \equiv 0$ if $q \geq p$, and in this case there is probability 0 of staying forever among the nonnegative integers. If $q < p, A = 1$ gives the maximal solution $x_n = 1 - (q/p)^{n+1}$ (and $0 \leq A < 1$ gives exactly the solutions that are not maximal). Compare (7.8) and Example 8.9.

Now consider the system (8.26) with $U = S - \{i_0\}$ for an arbitrary single state i_0:

$$\begin{cases} x_i = \sum_{j \neq i_0} p_{ij} x_j, & i \neq i_0, \\ 0 \leq x_i \leq 1, & i \neq i_0. \end{cases} \tag{8.27}$$

There is always the trivial solution—the one for which $x_i \equiv 0$.

THEOREM 8.5
An irreducible chain is transient if and only if (8.27) *has a nontrivial solution.*

Proof. The probabilities

$$1 - f_{ii_0} = P_i[X_n \neq i_0, n \geq 1], \quad i \neq i_0, \tag{8.28}$$

are by Theorem 8.4 the maximal solution of (8.27). Therefore (8.27) has a nontrivial solution if and only if $f_{ii_0} < 1$ for some $i \neq i_0$. If the chain is persistent, this is impossible by Theorem 8.3(ii).

Suppose the chain is transient. Since

$$\begin{aligned} f_{i_0 i_0} &= P_{i_0}[X_1 = i_0 + \sum_{n=2}^{\infty} \sum_{i \neq i_0} P_{i_0}[X_1 = i, X_2 \neq i_0, \ldots, X_{n-1} \neq i_0, X_n = i_0] \\ &= p_{i_0 i_0} + \sum_{i \neq i_0} p_{i_0 i} f_{i i_0}, \end{aligned}$$

and since $f_{i_0 i_0} < 1$, it follows that $f_{ii_0} < 1$ for some $i \neq i_0$. ■

Since the equations in (8.27) are homogeneous, the issue is whether they have a solution that is nonnegative, nontrivial, and *bounded*. If they do, $0 \leq x_i \leq 1$ can be arranged by rescaling.†

†See Problem 8.9.

EXAMPLE 8.11

In the simplest of *queueing models* the state space is $\{0, 1, 2, \ldots\}$ and the transition matrix has the form

$$\begin{bmatrix} t_0 & t_1 & t_2 & 0 & 0 & 0 & \cdots \\ t_0 & t_1 & t_2 & 0 & 0 & 0 & \cdots \\ 0 & t_0 & t_1 & t_2 & 0 & 0 & \cdots \\ 0 & 0 & t_0 & t_1 & t_2 & 0 & \cdots \\ 0 & 0 & 0 & t_0 & t_1 & t_2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

If there are i customers in the queue and $i \geq 1$, the customer at the head of the queue is served and leaves, and then 0, 1, or 2 new customers arrive (probabilities t_0, t_1, t_2), which leaves a queue of length $i-1, i$, or $i+1$. If $i = 0$, no one is served, and the new customers bring the queue length to 0, 1, or 2. Assume that t_0 and t_2 are positive, so that the chain is irreducible.

For $i_0 = 0$ the system (8.27) is

$$\begin{aligned} x_1 &= t_1 x_1 + t_2 x_2, \\ x_k &= t_0 x_{k-1} + t_1 x_k + t_2 x_{k+1}, \quad k \geq 2. \end{aligned} \tag{8.29}$$

Since t_0, t_1, t_2 have the form $q(1-t), t, p(1-t)$ for appropriate p, q, t, the second line of (8.29) has the form $x_k = px_{k+1} + qx_{k-1}, k \geq 2$. Now the solution [A19] is $A + B(q/p)^k = A + B(t_0/t_2)^k$ if $t_0 \neq t_2 (p \neq q)$ and $A + Bk$ if $t_0 = t_2 (p = q)$, and A can be expressed in terms of B because of the first equation in (8.29). The result is

$$x_k = \begin{cases} B((t_0/t_2)^k - 1) & \text{if } t_0 \neq t_2, \\ Bk & \text{if } t_0 = t_2. \end{cases}$$

There is a nontrivial solution if $t_0 < t_2$ but not if $t_0 \geq t_2$.

If $t_0 < t_2$, the chain is thus transient, and the queue size goes to infinity with proability 1. If $t_0 \geq t_2$, the chain is persistent. For a nonempty queue the expected increase in queue length in one step is $t_2 - t_0$, and the queue goes out of control if and only if this is positive.

Stationary Distributions

Suppose that the chain has initial probabilities π_i satisfying

$$\sum_{i \in S} \pi_i p_{ij} = \pi_j, \quad j \in S. \tag{8.30}$$

It then follows by induction that

$$\sum_{i \in S} \pi_i p_{ij}^{(n)} = \pi_j, \quad j \in S, \quad n = 0, 1, 2, \ldots. \tag{8.31}$$

If π_i is the probability that $X_0 = i$, then the left side of (8.31) is the probability that $X_n = j$, and thus (8.30) implies that the probability of $[X_n = j]$ is the same for all n. A set of probabilities satisfying (8.30) is for this reason called a *stationary distribution*. The existence of such a distribution implies that the chain is very stable.

To discuss this requires the notion of periodicity. The state j has *period* t if $p_{jj}^{(n)} > 0$ implies that t divides n and if t is the largest integer with this property. In other words, the period of j is the greatest common divisor of the set of integers

$$[n\colon n \geq 1, p_{jj}^{(n)} > 0]. \tag{8.32}$$

If the chain is irreducible, then for each pair i and j there exist r and s for which $p_{ij}^{(r)}$ and $p_{ji}^{(s)}$ are positive, and of course

$$p_{ii}^{(r+s+n)} \geq p_{ij}^{(r)} p_{jj}^{(n)} p_{ji}^{(s)}. \tag{8.33}$$

Let t_i and t_j be the periods of i and j. Taking $n = 0$ in this inequality shows that t_i divides $r + s$; and now it follows by the inequality that $p_{jj}^{(n)} > 0$ implies that t_i divides $r + s + n$ and hence divides n. Thus t_i divides each integer in the set (8.32), and so $t_i \leq t_j$. Since i and j can be interchanged in this argument, i and j have the same period. One can thus speak of the period of the chain itself in the irreducible case. The random walk on the line has period 2, for example. If the period is 1, the chain is *aperiodic*

Lemma 2. *In an irreducible, aperiodic chain, for each i and j,* $p_{ij}^{(n)} > 0$ *for all n exceeding some* $n_0(i, j)$

Proof. Since $p_{jj}^{(m+n)} \geq p_{jj}^{(m)} p_{jj}^{(n)}$, if M is the set (8.32) then $m \in M$ and $n \in M$ together imply $m + n \in M$. But it is a fact of number theory [A21] that if a set of positive integers is closed under addition and has greatest common divisor 1, then it contains all integers exceeding some n_1. Given i and j, choose r so that $p_{ij}^{(r)} > 0$. If $n > n_0 = n_1 + r$, then $p_{ij}^{(n)} \geq p_{ij}^{(r)} p_{jj}^{(n-r)} > 0$. ■

THEOREM 8.6

Suppose of an irreducible, aperiodic chain that there exists a stationary distribution—a solution of (8.30) *satisfying* $\pi_i \geq 0$ *and* $\sum_i \pi_i = 1$. *Then the chain is persistent,*

$$\lim_n p_{ij}^{(n)} = \pi_j \tag{8.34}$$

for all i and j, the π_j *are all positive, and the stationary distribution is unique.*

The main point of the conclusion is that the effect of the initial state wears off. Whatever the actual initial distribution $\{\alpha_i\}$ of the chain may be, if (8.34) holds, then it follows by the M-test that the probability $\sum_i \alpha_i p_{ij}^{(n)}$ of $[X_n = j]$ converges to π_j.

Proof. If the chain is transient, then $p_{ij}^{(n)} \to 0$ for all i and j by Theorem 8.3, and it follows by (8.31) and the M-test that π_j is identically 0, which contradicts $\sum_i \pi_i = 1$. The existence of a stationary distribution therefore implies that the chain is persistent.

Consider now a Markov chain with state space $S \times S$ and transition probabilities $p(ij, kl) = p_{ik}p_{jl}$ (it is easy to verify that these form a stochastic matrix). Call this the *coupled* chain; it describes the joint behavior of a pair of independent systems, each evolving according to the laws of the original Markov chain. By Theorem 8.1 there exists a Markov chain $(X_n, Y_n), n = 0, 1, \ldots$, having positive initial probabilities and transition probabilities

$$P[(X_{n+1}, Y_{n+1}) = (k, l) | (X_n, Y_n) = (i, j)] = p(ij, kl).$$

For n exceeding some n_0 depending on i, j, k, l, the probability $p^{(n)}(ij, kl) = p_{ik}^{(n)} p_{jl}^{(n)}$ is positive by Lemma 2. Therefore, the coupled chain is *irreducible*. (This proof that the coupled chain is irreducible requires only the assumptions that the original chain is irreducible and aperiodic, a fact needed again in the proof of Theorem 8.7.)

It is easy to check that $\pi(ij) = \pi_i \pi_j$ forms a set of stationary initial probabilities for the coupled chain, which, like the original one, must therefore be *persistent*. It follows that, for an arbitrary initial state (i, j) for the chain $\{(X_n, Y_n)\}$ and an arbitrary i_0 in S, one has $P_{ij}[(X_n, Y_n) = (i_0, i_0) \text{ i.o.}] = 1$. If τ is the smallest integer such that $X_\tau = Y_\tau = i_0$, then τ is finite with probability 1 under P_{ij}. The idea of the proof is now this: X_n starts in i and Y_n starts in j; once $X_n = Y_n = i_0$ occurs, X_n and Y_n follow identical probability laws, and hence the initial states i and j will lose their influence.

By (8.13) applied to the coupled chain, if $m \le n$, then

$$\begin{aligned} P_{ij}[(X_n, Y_n) = (k, l), \tau = m] \\ &= P_{ij}[(X_i, Y_t) \ne (i_0, i_0), t < m, (X_m, Y_m) = (i_0, i_0)] \\ &\times P_{i_0 i_0}[(X_{n-m}, Y_{n-m}) = (k, l)] \\ &= P_{ij}[\tau = m] p_{i_0 k}^{(n-m)} p_{i_0 l}^{(n-m)}. \end{aligned}$$

Adding out l gives $P_{ij}[X_n = k, \tau = m] = P_{ij}[\tau = m] p_{i_0 k}^{(n-m)}$, and adding out k gives $P_{ij}[Y_n = l, \tau = m] = P_{ij}[\tau = m] p_{i_0 l}^{(n-m)}$. Take $k = l$, equate probabilities,

and add over $m = 1, \ldots, n$:

$$P_{ij}[X_n = k, \tau \le n] = P_{ij}[Y_n = k, \tau \le n].$$

From this follows

$$\begin{aligned} P_{ij}[X_n = k] &\le P_{ij}[X_n = k, \tau \le n] + P_{ij}[\tau > n] \\ &= P_{ij}[Y_n = k, \tau \le n] + P_{ij}[\tau > n] \\ &\le P_{ij}[Y_n = k] + P_{ij}[\tau > n]. \end{aligned}$$

This and the same inequality with X and Y interchanged give

$$|p_{ik}^{(n)} - p_{jk}^{(n)}| = |P_{ij}[X_n = k] - P_{ij}[Y_n = k]| \le P_{ij}[\tau > n].$$

Since τ is finite with probability 1,

$$\lim_n |p_{ik}^{(n)} - p_{jk}^{(n)}| = 0. \tag{8.35}$$

(This proof of (8.35) goes through as long as the coupled chain is irreducible and persistent—no assumptions on the original chain are needed. This fact is used in the proof of the next theorem.)

By (8.31), $\pi_k - p_{jk}^{(n)} = \sum_i \pi_i (p_{ik}^{(n)} - p_{jk}^{(n)})$, and this goes to 0 by the M-test if (8.35) holds. Thus $\lim_n p_{jk}^{(n)} = \pi_k$. As this holds for each stationary distribution, there can be only one of them.

It remains to show that the π_j are all strictly positive. Choose r and s so that $p_{ij}^{(r)}$ and $p_{ji}^{(s)}$ are positive. Letting $n \to \infty$ in (8.33) shows that π_i is positive if π_j is; since some π_j is positive (they add to 1), all the π_i must be positive. ■

EXAMPLE 8.12

For the queueing model in Example 8.11 the equations (8.30) are

$$\begin{aligned} \pi_0 &= \pi_0\ t_0 + \pi_1 t_0, \\ \pi_1 &= \pi_0\ t_1 + \pi_1 t_1 + \pi_2 t_0, \\ \pi_2 &= \pi_0\ t_2 + \pi_1 t_2 + \pi_2 t_1 + \pi_3 t_0, \\ \pi_k &= \pi_{k-1} t_2 + \pi_k t_1 + \pi_{k+1} t_0, \quad k \ge 3. \end{aligned}$$

Again write t_0, t_1, t_2, as $q(1-t), t, p(1-t)$. Since the last equation here is $\pi_k = q\pi_{k+1} + p\pi_{k-1}$, the solution is

$$\pi_k = \begin{cases} A + B(p/q)^k = A + B(t_2/t_0)^k & \text{if } t_0 \ne t_2, \\ A + Bk & \text{if } t_0 = t_2 \end{cases}$$

for $k \geq 2$. If $t_0 < t_2$ and $\sum \pi_k$ converges, then $\pi_k \equiv 0$, and hence there is no stationary distribution; but this is not new, because it was shown in Example 8.11 that the chain is transient in this case. If $t_0 = t_2$, there is again no stationary distribution, and this is new because the chain was in Example 8.11 shown to be persistent in this case.

If $t_0 > t_2$, then $\sum \pi_k$ converges, provided $A = 0$. Solving for π_0 and π_1 in the first two equations of the system above gives $\pi_0 = Bt_2$ and $\pi_1 = Bt_2(1 - t_0)/t_0$. From $\sum_k \pi_k = 1$ it now follows that $B = (t_0 - t_2)/t_2$, and the π_k can be written down explicitly. Since $\pi_k = B(t_2/t_0)^k$ for $k \geq 2$, there is small chance of a large queue length.

If $t_0 = t_2$ in this queueing model, the chain is persistent (Example 8.11) but has no stationary distribution (Example 8.12). The next theorem describes the asymptotic behavior of the $p_{ij}^{(n)}$ in this case.

THEOREM 8.7

If an irreducible, aperiodic chain has no stationary distribution, then

$$\lim_n p_{ij}^{(n)} = 0 \tag{8.36}$$

for all i and j.

If the chain is transient, (8.36) follows from Theorem 8.3. What is interesting here is the persistent case.

Proof. By the argument in the proof of Theorem 8.6, the coupled chain is irreducible. If it is transient, then $\sum_n (p_{ij}^{(n)})^2$ converges by Theorem 8.2, and the conclusion follows.

Suppose, on the other hand, that the coupled chain is (irreducible and) persistent. Then the stopping-time argument leading to (8.35) goes through as before. If the $p_{ij}^{(n)}$ do not all go to 0, then there is an increasing sequence $\{n_u\}$ of integers along which some $p_{ij}^{(n)}$ is bounded away from 0. By the diagonal method [A14], it is possible by passing to a subsequence of $\{n_u\}$ to ensure that each $p_{ij}^{(n_u)}$ converges to a limit, which by (8.35) must be independent of i. Therefore, there is a sequence $\{n_u\}$ such that $\lim_u p_{ij}^{(n_u)} = t_j$ exists for all i and j, where t_j is nonnegative for all j and positive for some j. If M is a finite set of states, then $\sum_{j \in M} t_j = \lim_u \sum_{j \in M} p_{ij}^{(n_u)} \leq 1$, and hence $0 < t = \sum_j t_j \leq 1$. Now $\sum_{k \in M} p_{ik}^{(n_u)} p_{kj} \leq p_{ij}^{(n_u+1)} = \sum_k p_{ik} p_{kj}^{(n_u)}$; it is possible to pass to the limit ($u \to \infty$) inside the first sum (if M is finite) and inside the second sum (by the M-test), and hence $\sum_{k \in M} t_k p_{kj} \leq \sum_k p_{ik} t_j = t_j$. Therefore, $\sum_k t_k p_{kj} \leq t_j$; if one of these inequalities were strict, it would follow that $\sum_k t_k = \sum_j \sum_k t_k p_{kj} <$

$\sum_j t_j$, which is impossible. Therefore $\sum_k t_k p_{kj} = t_j$ for all j, and the ratios $\pi_j = t_j/t$ give a stationary distribution, contrary to the hypothesis. ■

The limits in (8.34) and (8.36) can be described in terms of mean return times. Let

$$\mu_j = \sum_{n=1}^{\infty} n f_{jj}^{(n)}; \tag{8.37}$$

if the series diverges, write $\mu_j = \infty$. In the persistent case, this sum is to be thought of as the average number of steps to first return to j, given that $X_0 = j$.†

Lemma 3. *Suppose that j is persistent and that $\lim_n p_{jj}^{(n)} = u$. Then $u > 0$ if and only if $\mu_j < \infty$, in which case $u = 1/\mu_j$.*

Under the convention that $0 = 1/\infty$, the case $u = 0$ and $\mu_j = \infty$ is consistent with the equation $u = 1/\mu_j$.

Proof. For $k \geq 0$ let $\rho_k = \sum_{n>k} f_{jj}^{(n)}$; the notation does not show the dependence on j, which is fixed. Consider the double series

$$\begin{aligned} f_{jj}^{(1)} + f_{jj}^{(2)} + f_{jj}^{(3)} + \cdots \\ + f_{jj}^{(2)} + f_{jj}^{(3)} + \cdots \\ + f_{jj}^{(3)} + \cdots \\ + \cdots . \end{aligned}$$

The kth row sums to $\rho_k (k \geq 0)$ and the nth column sums to $n f_{jj}^{(n)} (n \geq 1)$, and so [A27] the series in (8.37) converges if and only if $\sum_k \rho_k$ does, in which case

$$\mu_j = \sum_{k=0}^{\infty} \rho_k. \tag{8.38}$$

Since j is persistent, the P_j-probability that the system does not hit j up to time n is the probability that it hits j after time n, and this is ρ_n. Therefore,

$$\begin{aligned} 1 - P_{jj}^{(n)} &= P_j[X_n \neq j] \\ &= P_j[X_1 \neq j, \ldots, X_n \neq j] + \sum_{k=1}^{n-1} P_j[X_k = j, X_{k+1} \neq j, \ldots, X_n \neq j] \end{aligned}$$

†Since in general there is no upper bound to the number of steps to first return, it is not a simple random variable. It does come under the general theory in Chapter 4, and its expected value is indeed μ_j (and (8.38) is just (5.29)), but for the present the interpretation of μ_j as an average is informal. See Problem 23.11.

$$= \rho_n + \sum_{k=1}^{n-1} p_{jj}^{(k)} \rho_{n-k},$$

and since $\rho_0 = 1$,

$$1 = \rho_0 p_{jj}^{(n)} + \rho_1 p_{jj}^{(n-1)} + \cdots + \rho_{n-1} p_{jj}^{(1)} + \rho_n p_{jj}^{(0)}.$$

Keep only the first $k+1$ terms on the right here, and let $n \to \infty$; the result is $1 \geq (\rho_0 + \cdots + \rho_k)u$. Therefore $u > 0$ implies that $\sum_k \rho_k$ converges, so that $\mu_j < \infty$.

Write $x_{nk} = \rho_k p_{jj}^{(n-k)}$ for $0 \leq k \leq n$ and $x_{nk} = 0$ for $n < k$. Then $0 \leq x_{nk} \leq \rho_k$ and $\lim_n x_{nk} = \rho_k u$. If $\mu_j < \infty$, then $\sum_k \rho_k$ converges and it follows by the M-test that $1 = \sum_{k=0}^{\infty} x_{nk} \to \sum_{k=0}^{\infty} \rho_k u$. By (8.38), $1 = \mu_j u$, so that $u > 0$ and $u = 1/\mu_j$. ■

The law of large numbers bears on the relation $u = 1/\mu_j$ in the persistent case. Let V_n be the number of visits to state j up to time n. If the time from one visit to the next is about μ_j, then V_n should be about n/μ_j: $V_n/n \approx 1/\mu_j$. But (if $X_0 = j$) V_n/n has expected value $n^{-1} \sum_{k=1}^{n} p_{jj}^{(k)}$, which goes to u under the hypothesis of Lemma 3 [A30].

Consider an irreducible, aperiodic, persistent chain. There are two possibilities. If there is a stationary distribution, then the limits (8.34) are positive, and the chain is called *positive persistent*. It then follows by Lemma 3 that $\mu_j < \infty$ and $\pi_j = 1/\mu_j$ for all j. In this case, it is not actually necessary to assume persistence, since this follows from the existence of a stationary distribution. On the other hand, if the chain has no stationary distribution, then the limits (8.36) are all 0, and the chain is called *null persistent*. It then follows by Lemma 3 that $\mu_j = \infty$ for all j. This, taken together with Theorem 8.3, provides a complete classification:

THEOREM 8.8

For an irreducible, aperiodic chain there are three possibilities:

(i) *The chain is transient; then for all i and j, $\lim_n p_{ij}^{(n)} = 0$ and in fact $\sum_n p_{ij}^{(n)} < \infty$.*

(ii) *The chain is persistent but there exists no stationary distribution (the null persistent case); then for all i and j, $p_{ij}^{(n)}$ goes to 0 but so slowly that $\sum_n p_{ij}^{(n)} = \infty$, and $\mu_j = \infty$.*

(iii) *There exist stationary probabilities π_j and (hence) the chain is persistent (the positive persistent case); then for all i and j, $\lim_n p_{ij}^{(n)} = \pi_j > 0$ and $\mu_j = 1/\pi_j < \infty$.*

Since the asymptotic properties of the $p_{ij}^{(n)}$ are distinct in the three cases, these asymptotic properties in fact characterize the three cases.

EXAMPLE 8.13

Suppose that the states are 0, 1, 2, ... and the transition matrix is

$$\begin{bmatrix} q_0 & p_0 & 0 & 0 & \dots \\ q_1 & 0 & p_1 & 0 & \dots \\ q_2 & 0 & 0 & p_2 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

where p_i and q_i are positive. The state i represents the length of a success run, the conditional chance of a further success being p_i. Clearly the chain is irreducible and aperiodic.

A solution of the system (8.27) for testing for transience (with $i_0 = 0$) must have the form $x_k = x_1/p_1 \cdots p_{k-1}$. Hence there is a bounded, nontrivial solution, and the chain is transient, if and only if the limit α of $p_0 \cdots p_n$ is positive. But the chance of no return to 0 (for initial state 0) in n steps is clearly $p_0 \cdots p_{n-1}$; hence $f_{00} = 1 - \alpha$, which checks: the chain is persistent if and only if $\alpha = 0$.

Every solution of the steady-state equations (8.30) has the form $\pi_k = \pi_0 p_0 \cdots p_{k-1}$. Hence there is a stationary distribution if and only if $\sum_k p_0 \cdots p_k$ converges; this is the positive persistent case. The null persistent case is that in which $p_0 \cdots p_k \to 0$ but $\sum_k p_0 \cdots p_k$ diverges (which happens, for example, if $q_k = 1/k$ for $k > 1$).

Since the chance of no return to 0 in n steps is $p_0 \cdots p_{n-1}$, in the persistent case (8.38) gives $\mu_0 = \sum_{k=0}^{\infty} p_0 \cdots p_{k-1}$. In the null persistent case this checks with $\mu_0 = \infty$; in the positive persistent case it gives $\mu_0 = \sum_{k=0}^{\infty} \pi_k/\pi_0 = 1/\pi_0$, which again is consistent.

EXAMPLE 8.14

Since $\sum_j p_{ij}^{(n)} = 1$, possibilities (i) and (ii) in Theorem 8.8 are impossible in the finite case: A finite, irreducible, aperiodic Markov chain has a stationary distribution.

Exponential Convergence†

In the finite case, $p_{ij}^{(n)}$ converges to π_j at an exponential rate:

†This topic may be omitted.

THEOREM 8.9

If the state space is finite and the chain is irreducible and aperiodic, then there is a stationary distribution $\{\pi_i\}$, and

$$|p_{ij}^{(n)} - \pi_j| \le A\rho^n,$$

where $A \ge 0$ and $0 \le \rho < 1$.

Proof.† Let $m_j^{(n)} = \min_i p_{ij}^{(n)}$ and $M_j^{(n)} = \max_i p_{ij}^{(n)}$. By (8.10),

$$m_j^{(n+1)} = \min_i \sum_v p_{iv} p_{vj}^{(n)} \ge \min_i \sum_v p_{iv} m_j^{(n)} = m_j^{(n)},$$

$$M_j^{(n+1)} = \max_i \sum_v p_{iv} p_{vj}^{(n)} \le \max_i \sum_v p_{iv} M_j^{(n)} = M_j^{(n)},$$

Since obviously $m_j^{(n)} \le M_j^{(n)}$,

$$0 \le m_j^{(1)} \le m_j^{(2)} \le \cdots \le M_j^{(2)} \le M_j^{(1)} \le 1. \tag{8.39}$$

Suppose temporarily that all the p_{ij} are positive. Let s be the number of states and let $\delta = \min_{ij} p_{ij}$. From $\sum_j p_{ij} \ge s\delta$ follows $0 < \delta \le s^{-1}$. Fix states u and v for the moment; let $\sum'$ denote the summation over j in S satisfying $p_{uj} \ge p_{vj}$ and let $\sum''$ denote summation over j satisfying $p_{uj} < p_{vj}$. Then

$$\sum{}'(p_{uj} - p_{vj}) + \sum{}''(p_{uj} - p_{vj}) = 1 - 1 = 0. \tag{8.40}$$

Since $\sum' p_{vj} + \sum'' p_{uj} \ge s\delta$.

$$\sum{}'(p_{uj} - p_{vj}) = 1 - \sum{}'' P_{uj} - \sum{}' p_{vj} \le 1 - s\delta. \tag{8.41}$$

Apply (8.40) and then (8.41):

$$\begin{aligned} p_{uk}^{(n+1)} - p_{vk}^{(n+1)} &= \sum_j (p_{uj} - p_{vj}) p_{jk}^{(n)} \\ &\le \sum{}'(p_{uj} - p_{vj}) M_k^{(n)} + \sum{}''(p_{uj} - p_{vj}) m_k^{(n)} \\ &= \sum{}'(\rho_{uj} - p_{vj})(M_k^{(n)} - m_k^{(n)}) \\ &\le (1 - s\delta)(M_k^{(n)} - m_k^{(n)}). \end{aligned}$$

†For other proofs, see Problems 8.18 and 8.27.

Since u and v are arbitrary,

$$M_k^{(n+1)} - m_k^{(n+1)} \leq (1 - s\delta)(M_k^{(n)} - m_k^{(n)}).$$

Therefore, $M_k^{(n)} - m_k^{(n)} \leq (1 - s\delta)^n$. It follows by (8.39) that $m_j^{(n)}$ and $M_j^{(n)}$ have a common limit π_j and that

$$|p_{ij}^{(n)} - \pi_j| \leq (1 - s\delta)^n. \tag{8.42}$$

Take $A = 1$ and $\rho = 1 - \mathrm{s}\delta$. Passing to the limit in $\sum_i p_{vi}^{(n)} p_{ij} = p_{vj}^{(n+1)}$ shows that the π_i are stationary probabilities. (Note that the proof thus far makes almost no use of the preceding theory.)

If the p_{ij} are not all positive, apply Lemma 2: Since there are only finitely many states, there exists an m such that $p_{ij}^{(m)} > 0$ for all i and j. By the case just treated, $M_j^{(mt)} - m_j^{(mt)} \leq \rho^t$. Take $A = \rho^{-1}$ and then replace ρ by $\rho^{1/m}$. ■

EXAMPLE 8.15
Suppose that

$$P = \begin{bmatrix} p_0 & p_1 & \cdots & p_{s-1} \\ p_{s-1} & p_0 & \cdots & p_{s-2} \\ \cdots & \cdots & \cdots & \cdots \\ p_1 & p_2 & \cdots & p_0 \end{bmatrix}.$$

The rows of P are the cyclic permutations of the first row: $p_{ij} = p_{j-i}, j - i$ reduced modulo s. Since the columns of P add to 1 as well as the rows, the steady-state equations (8.30) have the solution $\pi_i \equiv s^{-1}$. If the p_i are all positive, the theorem implies that $p_{ij}^{(n)}$ converges to s^{-1} at an exponential rate. If $X_0, Y_1, Y_2, \ldots$ are independent random variables with range $\{0, 1, \ldots, s-1\}$, if each Y_n has distribution $\{p_0, \ldots, p_{s-1}\}$, and if $X_n = X_0 + Y_1 + \cdots + Y_n$, where the sum is reduced modulo s, then $P[X_n = j] \to s^{-1}$. The X_n describe a random walk on a circle of points, and whatever the initial distribution, the positions become equally likely in the limit.

Optimal Stopping†

Assume throughout the rest of the section that *S is finite*. Consider a function τ on Ω for which $\tau(\omega)$ is a nonnegative integer for each ω. Let $\mathscr{F}_n =$

†This topic may be omitted.

$\sigma(X_0, X_1, \ldots, X_n)$; τ is a *stopping time* or a *Markov time* if

$$[\omega\colon \tau(\omega) = n] \in \mathscr{F}_n \tag{8.43}$$

for $n = 0, 1, \ldots$. This is analogous to the condition (7.18) on the gambler's stopping time. It will be necessary to allow $\tau(\omega)$ to assume the special value ∞, but only on a set of probability 0. This has no effect on the requirement (8.43), which concerns finite n only.

If f is a real function on the state space, then $f(X_0), f(X_1), \ldots$ are simple random variables. Imagine an observer who follows the successive states $X_0, X_1, \ldots$ of the system. He stops at time τ, when the state is X_τ (or $X_{\tau(\omega)}(\omega)$), and receives an reward or payoff $f(X_\tau)$. The condition (8.43) prevents prevision on the part of the observer. This is a kind of game, the stopping time is a strategy, and the problem is to find a strategy that maximizes the expected payoff $E[f(X_\tau)]$. The problem in Example 8.5 had this form; there $S = \{1, 2, \ldots, r+1\}$, and the payoff function is $f(i) = i/r$ for $i \leq r$ (set $f(r+1) = 0$).

If $P(A) > 0$ and $Y = \sum_j y_j I_{B_j}$ is a simple random variable, the B_j forming a finite decomposition of Ω into $\mathscr{F}$-sets, the conditional expected value of Y given A is defined by

$$E[Y|A] = \sum y_i P(B_j|A).$$

Denote by E_i conditional expected values for the case $A = [X_0 = i]$:

$$E_i[Y] = E[Y|X_0 = i] = \sum_j y_j P_i(B_j).$$

The stopping-time problem is to choose τ so as to maximize simultaneously $E_i[f(X_\tau)]$ for all initial states i. If x lies in the range of f, which is finite, and if τ is everywhere finite, then $[\omega\colon f(X_{\tau(\omega)}(\omega)) = x] = \bigcup_{n=0}^{\infty}[\omega\colon \tau(\omega) = n, f(X_n(\omega)) = x]$ lies in $\mathscr{F}$, and so $f(X_\tau)$ is a simple random variable. In order that this always hold, put $f(X_{\tau(\omega)}(\omega)) = 0$, say, if $\tau(\omega) = \infty$ (which happens only on a set of probability 0).

The game with payoff function f has at i the *value*

$$v(i) = \sup E_i[f(X_\tau)]. \tag{8.44}$$

the supremum extending over all Markov times τ. It will turn out that the supremum here is achieved: there always exists an optimal stopping time. It will also turn out that there is an optimal τ that works for all initial states i. The problem is to calculate $v(i)$ and find the best τ. If the chain is irreducible, the system must pass through every state, and the best strategy is obviously to wait until the system enters a state for which f is maximal. This describes an

optimal τ, and $v(i) = \max f$ for all i. For this reason the interesting cases are those in which some states are transient and others are absorbing ($p_{ii} = 1$).

A function φ on S is *excessive* or *superharmonic*, if[†]

$$\varphi(i) \geq \sum_j p_{ij}, \varphi(J), \quad i \in S. \tag{8.45}$$

In terms of conditional expectation the requirement is $\varphi(i) \geq E_i[\varphi(X_1)]$.

Lemma 4. *The value function v is excessive.*

Proof. Given ϵ, choose for each j in S a "good" stopping time τ_j satisfying $E_j[f(X_{\tau_j})] > v(j) - \epsilon$. By (8.43), $[\tau_j = n] = [(X_0, \ldots, X_n) \in I_{jn}]$ for some set I_{jn} of $(n+1)$-long sequences of states. Set $\tau = n + 1 (n \geq 0)$ on the set $[X_1 = j] \cap [(X_1, \ldots, X_{n+1}) \in I_{jn}]$; that is, take one step and then from the new state X_1 add on the "good" stopping time for that state. Then τ is a stopping time and

$$\begin{aligned} E_i[f(X_\tau)] &= \sum_{n=0}^{\infty} \sum_j \sum_k P_i[X_1 = j, (X_1, \ldots, X_{n+1}) \in I_{jn}, X_{n+1} = k] f(k) \\ &= \sum_{n=0}^{\infty} \sum_j \sum_k p_{ij} p_j[(X_0, \ldots, X_n) \in I_{jn}, X_n = k] f(k) \\ &= \sum_j p_{ij} E_j[f(X_{\tau j})]. \end{aligned}$$

Therefore, $v(i) \geq E_i[f(X_\tau)] \geq \sum_j p_{ij}(v(j) - \epsilon) = \sum_j p_{ij} v(j) - \epsilon$. Since ϵ was arbitrary, v is excessive. ■

Lemma 5. *Suppose that φ is excessive.*

(i) *For all stopping times $\tau, \varphi(i) \geq E_i[\varphi(X_\tau)]$.*

(ii) *For all pairs of stopping times satisfying $\sigma \leq \tau, E_i[\varphi(X_\sigma)] \geq E_i[\varphi(X_\tau)]$.*

Part (i) says that for an excessive payoff function, $\tau \equiv 0$ represents an optimal strategy.

Proof. To prove (i), put $\tau_N = \min\{\tau, N\}$. Then τ_N is a stopping time, and

$$\begin{aligned} E_i[\varphi(X_{\tau N})] = &\sum_{n=o}^{N-1} \sum_k P_i[\tau = n, X_n = k] \varphi(k) \\ &+ \sum_k P_i[\tau \geq N, X_N = k] \varphi(k). \end{aligned} \tag{8.46}$$

[†]Compare the conditions (7.28) and (7.35).

Since $[\tau \geq N] = [\tau < N]^c \in F_{N-1}$, the final sum here is by (8.13)

$$\sum_k \sum_j P_i[\tau \geq N_i, X_{N-1} = j, X_N = k]\varphi(k)$$
$$= \sum_k \sum_j P_i[\tau \geq N, X_{N-1} = j]p_{jk}\varphi(k) \leq \sum_j P_i[\tau \geq N, X_{N-1} = j]\varphi(j).$$

Substituting this into (8.46) leads to $E_i[\varphi(X_{\tau_N})] \leq E_i[\varphi(X_{\tau_{N-1}})]$. Since $\tau_0 = 0$ and $E_i[\varphi(X_0)] = \varphi(i)$, it follows that $E_i[\varphi(X_{\tau_N})] \leq \varphi(i)$ for all N. But for $\tau(\omega)$ finite, $\varphi(X_{\tau_{N(\omega)}}(\omega)) \to \varphi(X_{\tau(\omega)}(\omega))$ (there is equality for large N), and so $E_i[\varphi(X_{\tau_N})] \to E_i[\varphi(X_\tau)]$ by Theorem 5.4.

The proof of (ii) is essentially the same. If $\tau_N = \min\{\tau, \sigma + N\}$, then τ_N is a stopping time, and

$$E_i[\varphi(X_{\tau_N})] = \sum_{m=0}^{\infty} \sum_{n=0}^{N-1} \sum_k P_i[\sigma = m, \tau = m + n, X_{m+n} = k]\varphi(k)$$
$$+ \sum_{m=0}^{\infty} \sum_k P_i[\sigma = m, \tau \geq m + N, X_{m+N} = k]\varphi(k).$$

Since $[\sigma = m, \tau \geq m + N] = [\sigma = m] - [\sigma = m, \tau < m + N] \in \mathscr{T}_{m+N-1}$, again $E_i[\varphi(X_{\tau_N})] \leq E_i[\varphi(X_{\tau_{N-1}})] \leq E_i[\varphi(X_{\tau_0})]$. Since $\tau_0 = \sigma$, part (ii) follows from part (i) by another passage to the limit. ■

Lemma 6. *If an excessive function φ dominates the payoff function f, then it dominates the value function v as well*

By definition, to say that g dominates h is to say that $g(i) \geq h(i)$ for all i.

Proof. By Lemma 5, $\varphi(i) \geq E_i[\varphi(X_\tau)] \geq E_i[f(X_\tau)]$ for all Markov times τ, and so $\varphi(i) \geq v(i)$ for all i. ■

Since $\tau \equiv 0$ is a stopping time, v dominates f. Lemmas 4 and 6 immediately characterize v:

THEOREM 8.10
The value function v is the minimal excessive function dominating f.

There remains the problem of constructing the optimal strategy τ. Let M be the set of states i for which $v(i) = f(i)$; M, the *support set*, is nonempty, since it at least contains those i that maximize f. Let $A = \cap_{n=0}^{\infty}[X_n \notin M]$ be the event that the system never enters M. The following argument shows that $P_i(A) = 0$ for each i. As this is trivial if $M = S$, assume that $M \neq S$. Choose $\delta > 0$ so that $f(i) \leq v(i) - \delta$ for $i \in S - M$. Now $E_i[f(X_\tau)] = \sum_{n=0}^{\infty} \sum_k P_i[\tau = n, X_n =$

$k]f(k)$; replacing the $f(k)$ by $v(k)$ or $v(k) - \delta$ according as $k \in M$ or $k \in S - M$ gives $E_i[f(X_\tau)] \le E_i[v(X_\tau)] - \delta P_i[X_\tau \in S - M] \le E_i[v(X_\tau)] - \delta P_i(A) \le v(i) - \delta P_i(A)$, the last inequality by Lemmas 4 and 5. Since this holds for every Markov time, taking the supremum over τ gives $P_i(A) = 0$. Whatever the initial state, the system is thus certain to enter the support set M.

Let $\tau_0(\omega) = \min[n\colon X_n(\omega) \in M]$ be the *hitting time* for M. Then τ_0 is a Markov time, and $\tau_0 = 0$ if $X_0 \in M$. It may be that $X_n(\omega) \notin M$ for all n, in which case $\tau_0(\omega) = \infty$, but as just shown, the probability of this is 0.

THEOREM 8.11

The hitting time τ_0 is optimal: $E_i[f(X_{\tau_0})] = v(i)$ for all i.

Proof. By the definition of $\tau_0, f(X_{\tau_0}) = v(X_{\tau_0})$. Put $\varphi(i) = E_i[f(X_{\tau_0})] = E_i[v(X_{\tau_0})]$. The first step is to show that φ is excessive. If $\tau_1 = \min[n\colon n \ge 1, X_n \in M]$, then τ_1 is a Markov time and

$$\begin{aligned} E_i[v(X_{\tau_1})] &= \sum_{n=1}^{\infty} \sum_{k \in M} P_i[X_1 \notin M, \ldots, X_{n-1} \notin M, X_n = k] v(k) \\ &= \sum_{n=1}^{\infty} \sum_{k \in M} \sum_{j \in S} p_{ij} P_j[X_0 \notin M, \ldots, X_{n-2} \notin M, X_{n-1} = k] v(k) \\ &= \sum_{j} p_{ij} E_j[v(X_{\tau_0})]. \end{aligned}$$

Since $\tau_0 \le \tau_1, E_i[v(X_{\tau_0})] \ge E_i[v(X_{\tau_1})]$ by Lemmas 4 and 5.

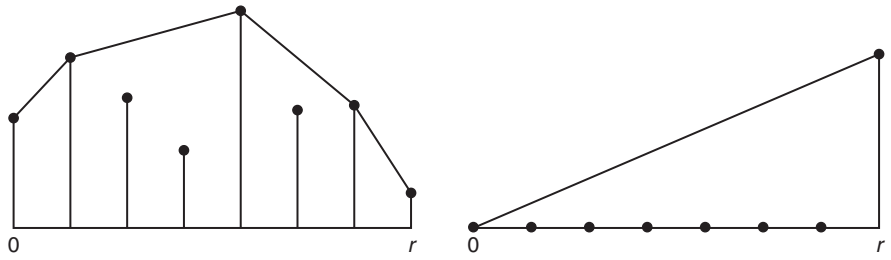
This shows that φ is excessive. And $\varphi(i) \le v(i)$ by the definition (8.44). If $\varphi(i) \ge f(i)$ is proved, it will follow by Theorem 8.10 that $\varphi(i) \ge v(i)$ and hence that $\varphi(i) = v(i)$. Since $\tau_0 = 0$ for $X_0 \in M$, if $i \in M$, then $\varphi(i) = E_i[f(X_0)] = f(i)$. Suppose that $\varphi(i) < f(i)$ for some values of i in $S - M$, and choose i_0 to maximize $f(i) - \varphi(i)$. Then $\psi(i) = \varphi(i) + f(i_0) - \varphi(i_0)$ dominates f and is excessive, being the sum of a constant and an excessive function. By Theorem 8.10, ψ must dominate v, so that $\psi(i_0) \ge v(i_0)$, or $f(i_0) \ge v(v_0)$. But this implies that $i_0 \in M$, a contraction. ■

The optimal strategy need not be unique. If f is constant, for example, all strategies have the same value.

EXAMPLE 8.16

For the symmetric random walk with absorbing barriers at 0 and r (Example 8.2) a function φ on $S = \{0, 1, \ldots, r\}$ is excessive if $\varphi(i) \ge \frac{1}{2}\varphi(i-1) + \frac{1}{2}\varphi(i+1)$ for $1 \le i \le r - 1$. The requirement is that φ give a concave function when extended by linear interpolation from S to the entire interval $[0, r]$. Hence v thus extended is the minimal concave function dominating f. The figure shows

the geometry: the ordinates of the dots are the values of f and the polygonal line describes v. The optimal strategy is to stop at a state for which the dot lies on the polygon.



If $f(r) = 1$ and $f(i) = 0$ for $i < r$, then v is a straight line; $v(i) = i/r$. The optimal Markov time τ_0 is the hitting time for $M = \{0, r\}$, and $v(i) = E_i[f(X_{\tau_0})]$ is the probability of absorption in the state r. This gives another solution of the gambler's ruin problem for the symmetric case.

EXAMPLE 8.17

For the selection problem in Example 8.5, the p_{ij} are given by (8.5) and (8.6) for $1 \le i \le r$, while $p_{r+1,r+1} = 1$. The payoff is $f(i) = i/r$ for $i \le r$ and $f(r+1) = 0$. Thus $v(r+1) = 0$, and since v is excessive,

$$v(i) \ge g(i) = \sum_{j=i+1}^{r} \frac{i}{j(j+1)} v(j), \quad 1 \le i < r. \tag{8.47}$$

By Theorem 8.10, v is the smallest function satisfying (8.47) and $v(i) \ge f(i) = i/r, 1 \le i \le r$. Since (8.47) puts no lower limit on $v(r)$, it follows that $v(r) = f(r) = 1$, and r lies in the support set M. By minimality,

$$v(i) = \max\{f(i), g(i)\}, \quad 1 \le i < r. \tag{8.48}$$

If $i \in M$, then $f(i) = v(i) \ge g(i) \ge \sum_{j=i+1}^{r} ij^{-1}(j-1)^{-1} f(j) = f(i) \sum_{j=i+1}^{r} (j-1)^{-1}$, and hence $\sum_{j=i+1}^{r} (j-1)^{-1} \le 1$. On the other hand, if this inequality holds and $i+1, \ldots, r$ all lie in M, then $g(i) = \sum_{j=i+1}^{r} ij^{-1}(j-1)^{-1} f(j) = f(i) \sum_{j=i+1}^{r} (j-1)^{-1} \le f(i)$, so that $i \in M$ by (8.48). Therefore, $M = \{i_r, i_r + 1, \ldots, r, r+1\}$, where i_r is determined by

$$\frac{1}{i_r} + \frac{1}{i_r + 1} + \cdots + \frac{1}{r-1} \le 1 < \frac{1}{i_r - 1} + \frac{1}{i_r} + \cdots + \frac{1}{r-1} \tag{8.49}$$

If $i < i_r$, so that $i \notin M$, then $v(i) > f(i)$ and so, by (8.48),

$$v(i) = g(i) = \sum_{j=i+1}^{i_r - 1} \frac{i}{j(j-1)} v(j) + \sum_{j=i_r}^{r} \frac{i}{j(j-1)} f(j)$$

$$= \sum_{j=i+1}^{i_r-1} \frac{i}{j(j-1)} v(j) + \frac{i}{r}\left(\frac{1}{i_r-1} + \cdots + \frac{1}{r-1}\right).$$

It follows by backward induction starting with $i = i_r - 1$ that

$$v(i) = p_r = \frac{i_r - 1}{r}\left(\frac{1}{i_r-1} + \cdots + \frac{1}{r-1}\right) \tag{8.50}$$

is constant for $1 \le i < i_r$.

In the selection problem as originally posed, $X_1 = 1$. The optimal strategy is to stop with the first X_n that lies in M. The princess should therefore reject the first $i_r - 1$ suitors and accept the next one who is preferable to all his predecessors (is dominant). The probability of success is p_r as given by (8.50). Failure can happen in two ways. Perhaps the first dominant suitor after i_r is not the best of all suitors; in this case the princess will be unaware of failure. Perhaps no dominant suitor comes after i_r; in this case the princess is obliged to take the last suitor of all and may be well aware of failure. Recall that the problem was to maximize the chance of getting the best suitor of all rather than, say, the chance of getting a suitor in the top half.

If r is large, (8.49) essentially requires that $\log r - \log i_r$ be near 1, so that $i_r \approx r/e$. In this case, $p_r \approx 1/e$.

Note that although the system starts in state 1 in the original problem, its resolution by means of the preceding theory requires consideration of all possible initial states.

This theory carries over in part to the case of infinite S, although this requires the general theory of expected values, since $f(X_\tau)$ may not be a simple random variable. Theorem 8.10 holds for infinite S if the payoff function is nonnegative and the value function is finite.[†] But then problems arise: Optimal strategies may not exist, and the probability of hitting the support set M may be less than 1. Even if this probability is 1, the strategy of stopping on first entering M may be the worst one of all.[‡]

PROBLEMS

8.1. Prove Theorem 8.1 for the case of finite S by constructing the appropriate probability measure on sequence space S^∞: Replace the summand on

[†]The only essential change in the argument is that Fatou's lemma (Theorem 16.3) must be used in place of Theorem 5.4 in the proof of Lemma 5.

[‡]See Problems 8.36 and 8.37.

the right in (2.21) by $\alpha_{u_1} p_{u_1 u_2}, \cdots p_{u_{n-1} u_n}$, and extend the arguments preceding Theorem 2.3. If $X_n(\cdot) = z_n(\cdot)$, then $X_1, X_2, \ldots$ is the appropriate Markov chain (here time is shifted by 1).

8.2. Let $Y_0, Y_1, \ldots$ be independent and identically distributed with $P[Y_n = 1] = p, P[Y_n = 0] = q = 1 - p, p \neq q$. Put $X_n = Y_n + Y_{n+1}$ (mode 2). Show that $X_0, X_1, \ldots$ is not a Markov chain even though $P[X_{n+1} = j | X_{n-1} = i] = P[X_{n+1} = j]$. Does this last relation hold for all Markov chains? Why?

8.3. Show by example that a function $f(X_0), f(X_1), \ldots$ of a Markov chain need not be a Markov chain.

8.4. Show that

$$f_{ij} \sum_{k=0}^{\infty} p_{jj}^{(k)} = \sum_{n=1}^{\infty} \sum_{m=1}^{n} f_{ij}^{m} p_{jj}^{(n-m)} = \sum_{n=1}^{\infty} p_{ij}^{(n)},$$

and prove that if j is transient, then $\sum_n p_{ij}^{(n)} < \infty$ for each i (compare Theorem 8.3(i)). If j is transient, then

$$f_{ij} = \sum_{n=1}^{\infty} p_{ij}^{(n)} \Big/ \left(1 + \sum_{n=1}^{\infty} p_{jj}^{(n)}\right).$$

Specialize to the case $i = j$: in addition to implying that i is transient (Theorem 8.2(i)), a finite value for $\sum_{n=1}^{\infty} p_{ii}^{(n)}$ suffices to determine f_{ii} exactly.

8.5. Call $\{x_i\}$ a *subsolution* of (8.24) if $x_i \leq \sum_j q_{ij} x_j$ and $0 \leq x_i \leq 1, i \in U$. Extending Lemma 1, show that a subsolution $\{x_i\}$ satisfies $x_i \leq \sigma_i$: The solution $\{\sigma_i\}$ of (8.24) dominates all subsolutions as well as all solutions. Show that if $x_i = \sum_j q_{ij} x_j$ and $-1 \leq x_i \leq 1$, then $\{|x_i|\}$ is a subsolution of (8.24).

8.6. Show by solving (8.27) that the unrestricted random walk on the line (Example 8.3) is persistent if and only if $p = \frac{1}{2}$.

8.7. **(a)** Generalize an argument in the proof of Theorem 8.5 to show that $f_{ik} = p_{ik} + \sum_{j \neq k} p_{ij} f_{jk}$. Generalize this further to

$$f_{ik} = f_{ik}^{(1)} + \cdots + f_{ik}^{(n)} + \sum_{j \neq k} P_i[X_1 \neq k, \ldots, X_{n-1} \neq k, X_n = j] f_{jk}.$$

(b) Take $k = i$. Show that $f_{ij} > 0$ if and only if $P_i[X_1 \neq i, \ldots, X_{n-1} \neq i, X_n = j] > 0$ for some n, and conclude that i is transient if and only if $f_{ji} < 1$ for some $j \neq i$ such that $f_{ij} > 0$.

(c) Show that an irreducible chain is transient if and only if for each i there is a $j \neq i$ such that $f_{ji} < 1$.

8.8. Suppose that $S = \{0, 1, 2, \ldots\}, p_{00} = 1$, and $f_{i0} > 0$ for all i.
(a) Show that $P_i(\cup_{j=1}^{\infty}[X_n = j \text{ i.o.}]) = 0$ for all i.
(b) Regard the state as the size of a population and interpret the conditions $p_{00} = 1$ and $f_{i0} > 0$ and the conclusion in part (a).

8.9. 8.5↑ Show for an irreducible chain that (8.27) has a nontrivial solution if and only if there exists a nontrivial, bounded sequence $\{x_i\}$ (not necessarily nonnegative) satisfying $x_i = \sum_{j \neq i_0} p_{ij} x_j, i \neq i_0$. (See the remark following the proof of Theorem 8.5.)

8.10. ↑ Show that an irreducible chain is transient if and only if (for arbitrary i_0) the system $y_i = \sum_j p_{ij} y_j, i \neq i_0$ (sum over *all* j), has a bounded, nonconstant solution $\{y_i, i \in S\}$.

8.11. Show that the P_i-probabilities of ever leaving U for $i \in U$ are the minimal solution of the system.

$$\begin{cases} z_i = \sum\limits_{j \in U} p_{ij} z_j + \sum\limits_{j \notin U} p_{ij}, & i \in U, \\ 0 \leq z_i \leq 1, & i \in U. \end{cases} \tag{8.51}$$

The constraint $z_i \leq 1$ can be dropped: the minimal solution automatically satisfies it, since $z_i \equiv 1$ is a solution.

8.12. Show that $\sup_{ij} n_0(i,j) = \infty$ is possible in Lemma 2.

8.13. Suppose that $\{\pi_i\}$ solves (8.30), where it is assumed that $\sum_i |\pi_i| < \infty$, so that the left side is well defined. Show in the irreducible case that the π_i are either all positive or all negative or all 0. Stationary probabilities thus exist in the irreducible case if and only if (8.30) has a nontrivial solution $\{\pi_i\}$ ($\sum_i \pi_i$ absolutely convergent).

8.14. Show by example that the coupled chain in the proof of Theorem 8.6 need not be irreducible if the original chain is not aperiodic.

8.15. Suppose that S consists of all the integers and

$$\begin{aligned} &p_{0,-1} = p_{0,0} = p_{0,+1} = \tfrac{1}{3}, \\ &p_{k,k-1} = q, \quad p_{k,k+1} = p, \qquad k \leq -1, \\ &p_{k,k-1} = p, \quad p_{k,k+1} = q, \qquad k \geq 1. \end{aligned}$$

Show that the chain is irreducible and aperiodic. For which p's is the chain persistent? For which p's are there stationary probabilities?

8.16. Show that the period of j is the greatest common divisor of the set

$$[n\colon n \geq 1, f_{ij}^{(n)} > 0]. \tag{8.52}$$

8.17. ↑ *Recurrent events*. Let $f_1, f_2, \ldots$ be nonnegative numbers with $f = \sum_{n=1}^{\infty} f_n \leq 1$. Define $u_1, u_2, \ldots$ recursively by $u_1 = f_1$ and

$$u_n = f_1 u_{n-1} + \cdots + f_{n-1} u_1 + f_n. \tag{8.53}$$

(a) Show that $f < 1$ if and only if $\sum_n u_n < \infty$.

(b) Assume that $f = 1$, set $\mu = \sum_{n=1}^{\infty} n f_n$, and assume that

$$\gcd[n\colon n \geq 1, f_n > 0] = 1. \tag{8.54}$$

Prove the *renewal theorem*: Under these assumptions, the limit $u = \lim_n u_n$ exists, and $u > 0$ if and only if $\mu < \infty$, in which case $u = 1/\mu$.

Although these definitions and facts are stated in purely analytical terms, they have a probabilistic interpretation: Imagine an event $\mathscr{E}$ that may occur at times $1, 2, \ldots$. Suppose f_n is the probability $\mathscr{E}$ occurs first at time n. Suppose further that at each occurrence of $\mathscr{E}$ the system starts anew, so that f_n is the probability that $\mathscr{E}$ next occurs n steps later. Such an $\mathscr{E}$ is called a *recurrent event*. If u_n is the probability that $\mathscr{E}$ occurs at time n, then (8.53) holds. The recurrent event $\mathscr{E}$ is called transient or persistent according as $f < 1$ or $f = 1$, it is called aperiodic if (8.54) holds, and if $f = 1$, μ is interpreted as the mean recurrence time.

8.18. **(a)** Let τ be the smallest integer for which $X_r = i_0$. Suppose that the state space is finite and that the p_{ij} are all positive. Find a ρ such that $\max_i (1 - p_{ii_0}) \leq \rho < 1$ and hence $P_i[\tau > n] \leq p^n$ for all i.

(b) Apply this to the coupled chain in the proof of Theorem 8.6: $|p_{ik}^{(n)} - p_{jk}^{(n)}| \leq p^n$. Now give a new proof of Theorem 8.9.

8.19. A thinker who owns r umbrellas wanders back and forth between home and office, taking along an umbrella (if there is one at hand) in rain (probability p) but not in shine (probability q). Let the state be the number of umbrellas at hand, irrespective of whether the thinker is at home or at work. Set up the transition matrix and find the stationary probabilities. Find the steady-state probability of his getting wet, and show that five umbrellas will protect him at the 5% level against any climate (any p).

8.20. **(a)** A transition matrix is *doubly stochastic* if $\sum_i p_{ij} = 1$ for each j. For a finite, irreducible, aperiodic chain with doubly stochastic transition matrix, show that the stationary probabilities are all equal.

(b) Generalize Example 8.15: Let S be a finite group, let $p(i)$ be probabilities, and put $p_{ij} = p(j \cdot i^{-1})$, where product and inverse refer to the group operation. Show that, if all $p(i)$ are positive, the states are all equally likely in the limit.

(c) Let S be the symmetric group on 52 elements. What has (b) to say about card shuffling?

8.21. A set C in S is *closed* if $\sum_{j\in C} p_{ij} = 1$ for $i \in C$: once the system enters C it cannot leave. Show that a chain is irreducible if and only if S has no proper closed subset.

8.22. ↑ Let T be the set of transient states and define persistent states i and j (if there are any) to be equivalent if $f_{ij} > 0$. Show that this is an equivalence relation on $S - T$ and decomposes it into equivalence classes $C_1, C_2, \ldots$, so that $S = T \cup C_1 \cup C_2 \cup \cdots$. Show that each C_m is closed and that $f_{ij} = 1$ for i and j in the same C_m.

8.23. 8.11 8.21 ↑ Let T be the set of transient states and let C be any closed set of persistent states. Show that the P_i-probabilities of eventual absorption in C for $i \in T$ are the minimal solution of

$$\begin{cases} y_i = \sum_{j\in T} p_{ij} y_i + \sum_{j\in C} p_{ij}, & i \in T, \\ 0 \le y_i \le 1, & i \in T. \end{cases} \tag{8.55}$$

8.24. Suppose that an irreducible chain has period $t > 1$. Show that S decomposes into sets $S_0, \ldots, S_{t-1}$ such that $p_{ij} > 0$ only if $i \in S_\nu$ and $j \in S_{\nu+1}$ for some ν ($\nu + 1$ reduced modulo t). Thus the system passes through the S_ν in cyclic succession.

8.25. ↑ Suppose that an irreducible chain of period $t > 1$ has a stationary distribution $\{\pi_j\}$. Show that, if $i \in S_\nu$ and $j \in S_{\nu+\alpha}$ ($\nu + \alpha$ reduced modulo t), then $\lim_n p_{ij}^{(nt+\alpha)} = \pi_j$. Show that $\lim_n n^{-1} \sum_{m=1}^{n} p_{ij}^{(m)} = \pi_j / t$ for all i and j.

8.26. *Eigenvalues*. Consider an irreducible, aperiodic chain with state space $\{1, \ldots, s\}$. Let $r_0 = (\pi_1, \ldots, \pi_s)$ be (Example 8.14) the row vector of stationary probabilities, and let c_0 be the column vector of 1's; then r_0 and c_0 are left and right eigenvectors of P for the eigenvalue $\lambda = 1$.

(a) Suppose that r is a left eigenvector for the (possibly complex) eigenvalue λ: $rP = \lambda r$. Prove: If $\lambda = 1$, then r is a scalar multiple of r_0 ($\lambda = 1$ has geometric multiplicity 1). If $\lambda \ne 1$, then $|\lambda| < 1$ and $rc_0 = 0$ (the 1×1 product of $1 \times s$ and $s \times 1$ matrices).

(b) Suppose that c is a right eigenvector: $Pc = \lambda c$. If $\lambda = 1$, then c is a scalar multiple of c_0 (again the geometric multiplicity is 1). If $\lambda \ne 1$, then again $|\lambda| < 1$, and $r_0 c = 0$.

8.27. ↑ Suppose P is diagonalizable; that is, suppose there is a nonsingular C such that $C^{-1}PC = \Lambda$, where Λ is a diagonal matrix. Let $\lambda_1, \ldots, \lambda_s$ be the diagonal elements of Λ, let $c_1, \ldots, c_s$ be the successive columns of C, let $R = C^{-1}$, and let $r_1, \ldots, r_s$ be the successive rows of R.

(a) Show that c_i and r_i are right and left eigenvectors for the eigenvalue $\lambda_i, i = 1, \ldots, s$. Show that $r_i c_j = \delta_{ij}$. Let $A_i = c_i r_i (s \times s)$. Show that Λ^n is a diagonal matrix with diagonal elements $\lambda_1^n, \ldots, \lambda_s^n$ and that $P^n = C\Lambda^n R = \sum_{u=1}^{s} \lambda_u^n A_u, n \geq 1$.

(b) Part (a) goes through under the sole assumption that P is a diagonalizable matrix. Now assume also that it is an irreducible, aperiodic stochastic matrix, and arrange the notation so that $\lambda_1 = 1$. Show that each row of A_1 is the vector $(\pi_1, \ldots, \pi_s)$ of stationary probabilities. Since

$$P^n = A_1 + \sum_{u=2}^{s} \lambda_u^n A_u \tag{8.56}$$

and $|\lambda_u| < 1$ for $2 \leq u \leq s$, this proves exponential convergence once more.

(c) Write out (8.56) explicitly for the case $s = 2$.

(d) Find an irreducible, aperiodic stochastic matrix that is not diagonalizable.

8.28. $\uparrow$

(a) Show that the eigenvalue $\lambda = 1$ has geometric multiplicity 1 if there is only one closed, irreducible set of states; there may be transient states, in which case the chain itself is not irreducible.

(b) Show, on the other hand, that if there is more than one closed, irreducible set of states, then $\lambda = 1$ has geometric multiplicity exceeding 1.

(c) Suppose that there is only one closed, irreducible set of states. Show that the chain has period exceeding 1 if and only if there is an eigenvalue other than 1 on the unit circle.

8.29. Suppose that $\{X_n\}$ is a Markov chain with state space S, and put $Y_n = (X_n, X_{n+1})$. Let T be the set of pairs (i, j) such that $p_{ij} > 0$ and show that $\{Y_n\}$ is a Markov chain with state space T. Write down the transition probabilities. Show that, if $\{X_n\}$ is irreducible and aperiodic, so is $\{Y_n\}$. Show that, if π_i are stationary probabilities for $\{X_n\}$, then $\pi_i p_{ij}$ are stationary probabilities for $\{Y_n\}$.

8.30. 6.10 8.29$\uparrow$ Suppose that the chain is finite, irreducible, and aperiodic and that the initial probabilities are the stationary ones. Fix a state i, let $A_n = [X_i = i]$, and let N_n be the number of passages through i in the first n steps. Calculate α_n and β_n as defined by (5.41). Show that $\beta_n - \alpha_n^2 = O(1/n)$, so that $n^{-1}N_n \to \pi_i$ with probability 1. Show for a function f on the state space that $n^{-1}\sum_{k=1}^{n} f(X_k) \to \sum_i \pi_i f(i)$ with probability 1. Show that $n^{-1}\sum_{k=1}^{n} g(X_k, X_{k+1}) \to \sum_{ij} \pi_{ij} p_{ij} g(i, j)$ for functions g on $S \times S$.

8.31. 6.14 8.30↑ If $X_0(\omega) = i_0, \ldots, X_n(\omega) = i_n$ for states $i_0, \ldots, i_n$, put $p_n(\omega) = \pi_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}$, so that $p_n(\omega)$ is the probability of the observation observed. Show that $-n^{-1} \log p_n(\omega) \to h = -\sum_{ij} \pi_i p_{ij} \log p_{ij}$ with probability 1 if the chain is finite, irreducible, and aperiodic. Extend to this case the notions of source, entropy, and asymptotic equipartition.

8.32. A sequence $\{X_n\}$ is a Markov chain *of second order* if $P[X_{n+1} = j | X_0 = i_0, \ldots, X_n = i_n] = P[X_{n+1} = j | X_{n-1} = i_{n-1}, X_n = i_n] = p_{i_{n-1} i_n; j}$. Show that nothing really new is involved because the sequence of pairs (X_n, X_{n+1}) is an ordinary Markov chain (of first order). Compare Problem 8.29. Generalize this idea into chains of order r.

8.33. Consider a chain on $S = \{0, 1, \ldots, r\}$, where 0 and r are absorbing states and $p_{i,i+1} = p_i > 0, p_{i,i-1} = q_i = 1 - p_i > 0$ for $0 < i < r$. Identify state i with a point z_i on the line, where $0 = z_0 < \cdots < z$, and the distance from z_i to z_{i+1} is q_i / p_i times that from z_{i-1} to z_i. Given a function φ on S, consider the associated function $\hat{\varphi}$ on $[0, z_r]$ defined at the z_i by $\hat{\varphi}(z_i) = \varphi(i)$ and in between by linear interpolation. Show that φ is excessive if and only if $\hat{\varphi}$ is concave. Show that the probability of absorption in r for initial state i is t_{i-1} / t_{r-1}, where $t_i = \sum_{k=0}^{i} q_1 \cdots q_k / p_1 \cdots p_k$. Deduce (7.7). Show that in the new scale the expected distance moved on each step is 0.

8.34. Suppose that a finite chain is irreducible and aperiodic. Show by Theorem 8.9 that an excessive function must be constant.

8.35. *A zero–one law*. Let the state space S contain s points, and suppose that $\epsilon_n = \sup_{ij} |p_{ij}^{(n)} - \pi_j| \to 0$, as holds under the hypotheses of Theorem 8.9. For $a \le b$, let $\mathscr{G}_a^b$ be the σ-field generated by the sets $[X_a = u_a, \ldots, X_b = u_b]$. Let $\mathscr{T}_a = \sigma(\bigcup_{b=a}^{\infty} \mathscr{G}_a^b)$ and $\mathscr{T} = \bigcap_{a=1}^{\infty} \mathscr{T}_a$. Show that $|P(A \cap B) - P(A)P(B)| \le s(\epsilon_n + \epsilon_{b+n})$ for $A \in \mathscr{G}_0^b$ and $B \in \mathscr{G}_{b+n}^{b+m}$; the ϵ_{b+n} can be suppressed if the initial probabilities are the stationary ones. Show that this holds for $A \in \mathscr{G}_0^b$ and $B \in \mathscr{T}_{b+n}$. Show that $C \in \mathscr{T}$ implies that $P(C)$ is either 0 or 1.

8.36.[†] Alter the chain in Example 8.13 so that $q_0 = 1 - p_0 = 1$ (the other p_i and q_i still positive). Let $\beta = \lim_n p_1 \cdots p_n$ and assume that $\beta > 0$. Define a payoff function by $f(0) = 1$ and $f(i) = 1 - f_{i0}$ for $i > 0$. If $X_0, \ldots, X_n$ are positive, put $\sigma_n = n$; otherwise let σ_n be the smallest k such that $X_k = 0$. Show that $E_i[f(X_{\sigma_n})] \to 1$ as $n \to \infty$, so that

[†]The final three problems in this section involve expected values for random variables with infinite range.

$v(i) \equiv 1$. Thus the support set is $M = \{0\}$, and for an initial state $i > 0$ the probability of ever hitting M is $f_{i0} < 1$.

For an arbitrary finite stopping time τ, choose n so that $P_i[\tau < n = \sigma_n] > 0$. Then $E_i[f(X_\tau)] \le 1 - f_{i+n,0}P_i[\tau < n = \sigma_n] < 1$. Thus no strategy achieves the value $v(i)$ (except of course for $i = 0$).

8.37. ↑ Let the chain be as in the preceding problem, but assume that $\beta = 0$, so that $f_{i0} = 1$ for all i. Suppose that $\lambda_1, \lambda_2, \ldots$ exceed 1 and that $\lambda_1 \cdots \lambda_n \to \lambda < \infty$; put $f(0) = 0$ and $f(i) = \lambda_1 \cdots \lambda_{i-1}/p_1 \cdots p_{i-1}$. For an arbitrary (finite) stopping time τ, the event $[\tau = n]$ must have the form $[(X_0, \ldots, X_n) \in I_n]$ for some set I_n of $(n+1)$-long sequences of states. Show that for each i there is at most one $n \ge 0$ such that $(i, i+1, \ldots, i+n) \in I_n$. If there is no such n, then $E_i[f(X_\tau)] = 0$. If there is one, then

$$E_i[f(X_\tau)] = P_i[(X_0, \ldots, X_n) = (i, \ldots, i+n)]f(i+n),$$

and hence the only possible values of $E_i[f(X_\tau)]$ are

$$0, f(i), p_i f(i+1) = f(i)\lambda_i, \quad p_i p_{i+1} f(i+2) = f(i)\lambda_i \lambda_{i+1}, \ldots.$$

Thus $v(i) = f(i)\lambda/\lambda_1 \cdots \lambda_{i-1}$ for $i \ge 1$; no strategy this value. The support set is $M = \{0\}$, and the hitting time τ_0 for M is finite, but $E_i[f(X_{\tau_0})] = 0$.

8.38. 5.12 ↑ Consider an irreducible, aperiodic, positive persistent chain. Let τ_j be the smallest n such that $X_n = j$, and let $m_{ij} = E_i[\tau_j]$. Show that there is an r such that $p = P_j[X_1 \ne j, \ldots, X_{r-1} \ne j, X_r = i]$ is positive; from $f_{jj}^{(n+r)} \ge p f_{ij}^{(n)}$ and $m_{jj} < \infty$, conclude that $m_{ij} < \infty$ and $m_{ij} = \sum_{n=0}^{\infty} P_i[\tau_j > n]$. Starting from $p_{ij}^{(t)} = \sum_{s=1}^{t} f_{ij}^{(s)} p_{jj}^{(t-s)}$, show that

$$\sum_{t=1}^{n} (p_{ij}^{(t)} - p_{jj}^{(t)}) = 1 - \sum_{m=0}^{n} p_{jj}^{(n-m)} P_i[\tau_j > m].$$

Use the M-test to show that

$$\pi_j m_{ij} = 1 + \sum_{n=1}^{\infty} (p_{jj}^{(n)} - p_{ij}^{(n)}).$$

If $i = j$, this gives $m_{jj} = 1/\pi_j$ again; if $i \ne j$, it shows how in principle m_{ij} can be calculated from the transition matrix and the stationary probabilities.

SECTION 9 LARGE DEVIATIONS AND THE LAW OF THE ITERATED LOGARITHM†

It is interesting in connection with the strong law of large numbers to estimate the rate at which S_n/n converges to the mean m. The proof of the strong law used upper bounds for the probabilities $P[|S_n - m| \geq \alpha]$ for large α. Accurate upper and lower bounds for these probabilities will lead to the law of the iterated logarithm, a theorem giving very precise rates for $S_n/n \to m$.

The first concern will be to estimate the probability of large deviations from the mean, which will require the method of moment generating functions. The estimates will be applied first to a problem in statistics and then to the law of the iterated logarithm.

Moment Generating Functions

Let X be a simple random variable assuming the distinct values $x_1, \ldots, x_1$ with respective probabilities $p_1, \ldots, p_1$. Its *moment generating function* is

$$M(t) = E[e^{tX}] = \sum_{i=1}^{1} p_i e^{tx_i}. \tag{9.1}$$

(See (5.19) for expected values of functions of random variables.) This function, defined for all real t, can be regarded as associated with X itself or as associated with its distribution—that is, with the measure on the line having mass p_i at x_i (see (5.12)).

If $c = \max_i |x_i|$, the partial sums of the series $e^{tX} = \sum_{k=0}^{\infty} t^k X^k/k!$ are bounded by $e^{|t|c}$, and so the corollary to Theorem 5.4 applies:

$$M(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k]. \tag{9.2}$$

Thus $M(t)$ has a Taylor expansion, and as follows from the general theory [A29], the coefficient of t^k must be $M^{(k)}(0)/k!$ Thus

$$E[X^k] = M^{(k)}(0). \tag{9.3}$$

Furthermore, term-by-term differentiation in (9.1) gives

$$M^{(k)}(t) = \sum_{i=1}^{1} p_i x_i^k e^{tx} = E[X^k e^{tX}];$$

†This section may be omitted.

taking $t = 0$ here gives (9.3) again. Thus the moments of X can be calculated by successive differentiation, whence $M(t)$ gets its name. Note that $M(0) = 1$.

EXAMPLE 9.1
If X assumes the values 1 and 0 with probabilities p and $q = 1 - p$, as in Bernoulli trials, its moment generating function is $M(t) = pe^t + q$. The first two moments are $M'(0) = p$ and $M''(0) = p$, and the variance is $p - p^2 = pq$.

If $X_1, \ldots, X_n$ are independent, then for each t (see the argument following (5.10)), $e^{tX_1}, \ldots, e^{tX_n}$ are also independent. Let M and $M_1, \ldots, M_n$ be the respective moment generating functions of $S = X_1 + \cdots + X_n$ and of $X_1, \ldots, X_n$; of course, $e^{tS} = \Pi_i e^{tX_i}$. Since by (5.25) expected values multiply for independent random variables, there results the fundamental relation

$$M(t) = M_1(t) \cdots M_n(t). \tag{9.4}$$

This is an effective way of calculating the moment generating function of the sum S. The real interest, however, centers on the distribution of S, and so it is important to know that distributions can in principle be recovered from their moment generating functions.

Consider along with (9.1) another finite exponential sum $N(t) = \sum_j q_j e^{ty_j}$, and suppose that $M(t) = N(t)$ for all t. If $x_{i_0} = \max x_i$ and $y_{j_0} = \max y_j$, then $M(t) \sim p_{i_0} e^{ix_{i_0}}$ and $N(t) \sim q_{j_0} e^{ty_{j_0}}$ as $t \to \infty$, and so $x_{i_0} = y_{j_0}$ and $p_{i_0} = q_{i_0}$. The same argument now applies to $\sum_{i \neq i_0} p_i e^{tx_i} = \sum_{j \neq j_0} q_j e^{ty_j}$, and it follows inductively that with appropriate relabeling, $x_i = y_i$ and $p_i = q_i$ for each i. Thus *the function* (9.1) *does uniquely determine the* x_i *and* p_i.

EXAMPLE 9.2
If $X_1, \ldots, X_n$ are independent, each assuming values 1 and 0 with probabilities p and q, then $S = X_1 + \cdots + X_n$ is the number of successes in n Bernoulli trials. By (9.4) and Example 9.1, S has the moment generating function

$$E[e^{tS}] = (pe^t + q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} e^{tk}.$$

The right-hand form shows this to be the moment generating function of a distribution with mass $\binom{n}{k} p^k q^{n-k}$ at the integer $k, 0 \leq k \leq n$. The uniqueness just established therefore yields the standard fact that $P[S = k] = \binom{n}{k} p^k q^{n-k}$.

The *cumulant generating function of* X (or of its distribution) is

$$C(t) = \log M(t) = \log E[e^{tX}]. \tag{9.5}$$

(Note that $M(t)$ is strictly positive.) Since $C' = M'/M$ and $C'' = (MM'' - (M')^2)/M^2$, and since $M(0) = 1$,

$$C(0) = 0, \quad C'(0) = E[X], \quad C''(0) = \mathrm{Var}[X]. \tag{9.6}$$

Let $m_k = E[X^k]$. The leading term in (9.2) is $m_0 = 1$, and so a formal expansion of the logarithm in (9.5) gives

$$C(t) = \sum_{v=1}^{\infty} \frac{(-1)^{v+1}}{v} \left(\sum_{k=1}^{\infty} \frac{m_k}{k!} t^k \right)^v. \tag{9.7}$$

Since $M(t) \to 1$ as $t \to 0$, this expression is valid for t in some neighborhood of 0. By the theory of series, the powers on the right can be expanded and terms with a common factor t^i collected together. This gives an expansion

$$C(t) = \sum_{i=1}^{\infty} \frac{c_i}{i!} t^i, \tag{9.8}$$

valid in some neighborhood of 0.

The c_i are the *cumulants* of X. Equating coefficients in the expansions (9.7) and (9.8) leads to $c_1 = m_1$ and $c_2 = m_2 - m_1^2$, which checks with (9.6). Each c_i can be expressed as a polynomial in $m_1, \ldots, m_i$ and conversely, although the calculations soon become tedious. If $E[X] = 0$, however, so that $m_1 = c_1 = 0$, it is not hard to check that

$$c_3 = m_3, \quad c_4 = m_4 - 3m_2^2. \tag{9.9}$$

Taking logarithms converts the multiplicative relation (9.4) into the additive relation

$$C(t) = C_1(t) + \cdots + C_n(t) \tag{9.10}$$

for the corresponding cumulant generating functions; it is valid in the presence of independence. By this and the definition (9.8), it follows that cumulants add for independent random variables.

Clearly, $M''(t) = E[X^2 e^{tX}] \geq 0$. Since $(M'(t))^2 = E^2[Xe^{tX}] \leq E[e^{tX}]$. $E[X^2 e^{tX}] = M(t)M''(t)$ by Schwarz's inequality (5.36), $C''(t) \geq 0$. Thus *the moment generating function and the cumulant generating function are both convex*.

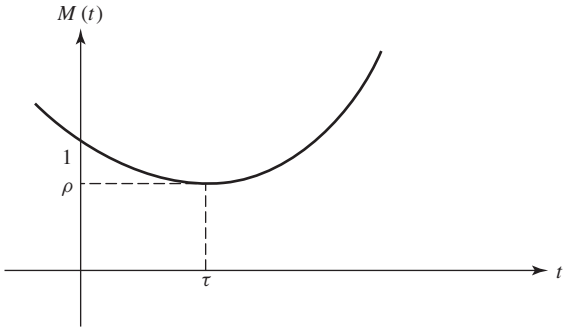
Large Deviations

Let Y be a simple random variable assuming values y_j with probabilities p_j. The problem is to estimate $P[Y \geq \alpha]$ when Y has mean 0 and α is positive. It is notationally convenient to subtract α away from Y and instead estimate $P[Y \geq 0]$ when Y has negative mean.

Assume then that

$$E[Y] < 0, \quad P[Y > 0] > 0, \tag{9.11}$$

the second assumption to avoid trivialities. Let $M(t) = \sum_j p_j e^{ty_j}$ be the moment generating function of Y. Then $M'(0) < 0$ by the first assumption in



(9.11), and $M(t) \to \infty$ as $t \to \infty$ by the second. Since $M(t)$ is convex, it has its minimum ρ at a positive argument τ:

$$\inf_t M(t) = M(\tau) = \rho, \qquad 0 < \rho < 1, \quad \tau > 0. \tag{9.12}$$

Construct (on an entirely irrelevant probability space) an auxiliary random variable Z such that

$$P[Z = y_j] = \frac{e^{\tau y_j}}{\rho} P[Y = y_j] \tag{9.13}$$

for each y_j in the range of Y. Note that the probabilities on the right do add to 1. The moment generating function of Z is

$$E[e^{iZ}] = \sum_j \frac{e^{\tau y_j}}{\rho} p_j e^{ty_j} = \frac{M(\tau + t)}{\rho}, \tag{9.14}$$

and therefore

$$E[Z] = \frac{M'(\tau)}{\rho} = 0, \quad s^2 = E[Z^2] = \frac{M''(\tau)}{\rho} > 0. \tag{9.15}$$

For all positive t, $P[Y \geq 0] = P[e^{tY} \geq 1] \leq M(t)$ by Markov's inequality (5.31), and hence

$$P[Y \geq 0] \leq \rho. \tag{9.16}$$

Inequalities in the other direction are harder to obtain. If $\sum'$ denotes summation over those indices j for which $y_j \geq 0$, then

$$P[Y \geq 0] = \sum{}' p_j = \rho \sum{}' e^{-\tau y_j} P[Z = y_j]. \tag{9.17}$$

Put the final sum here in the form $e^{-\theta}$, and let $p = P[Z \geq 0]$. By (9.16), $\theta \geq 0$. Since $\log x$ is concave, Jensen's inequality (5.33) gives

$$\begin{aligned} -\theta &= \log \sum{}' e^{-\tau y_j} p^{-1} P[Z = y_j] + \log p \\ &\geq \sum{}' (-\tau y_j) p^{-1} P[Z = y_j] + \log p \\ &= -\tau s p^{-1} \sum{}' \frac{y_j}{s} P[Z = y_j] + \log p. \end{aligned}$$

By (9.15) and Lyapounov's inequality (5.37),

$$\sum{}' \frac{y_j}{s} P[Z = y_j] \leq \frac{1}{s} E[|Z|] \leq \frac{1}{s} E^{1/2}[Z^2] = 1.$$

The last two inequalities give

$$0 \leq \theta \leq \frac{\tau s}{P[Z \geq 0]} - \log P[Z \geq 0]. \tag{9.18}$$

This proves the following result.

THEOREM 9.1
Suppose that Y satisfies (9.11)*. Define ρ and τ by* (9.12)*, let Z be a random variable with distribution* (9.13)*, and define s^2 by* (9.15)*. Then $P[Y \geq 0] = \rho e^{-\theta}$, where θ satisfies* (9.18)*.*

To use (9.18) requires a lower bound for $P[Z \geq 0]$.

THEOREM 9.2
If $E[Z] = 0, E[Z^2] = s^2$, and $E[Z^4] = \xi^4 > 0$, then $P[Z \geq 0] \geq s^4/4\xi^4$.[†]

[†]For a related result, see Problem 25.19.

Proof. Let $Z^+ = ZI_{[Z \geq 0]}$ and $Z^- = -ZI_{[Z<0]}$. Then Z^+ and Z^- are nonnegative, $Z = Z^+ - Z^-, Z^2 = (Z^+)^2 + (Z^-)^2$, and

$$s^2 = E[(Z^+)^2] + E[(Z^-)^2]. \tag{9.19}$$

Let $p = P[Z \geq 0]$. By Schwarz's inequality (5.36),

$$\begin{aligned} E[(Z^+)^2] &= E[I_{[Z \geq 0]}Z^2] \\ &\leq E^{1/2}[I^2_{[Z \geq 0]}]E^{1/2}[Z^4] = p^{1/2}\xi^2. \end{aligned}$$

By Hölder's inequality (5.35) (for $p = \frac{3}{2}$ and $q = 3$)

$$\begin{aligned} E[(Z^-)^2] &= E[(Z^-)^{2/3}(Z^-)^{4/3}] \\ &\leq E^{2/3}[Z^-]E^{1/3}[(Z^-)^4] \leq E^{2/3}[Z^-]\xi^{4/3}. \end{aligned}$$

Since $E[Z] = 0$, another application of Hölder's inequality (for $p = 4$ and $q = \frac{4}{3}$) gives

$$\begin{aligned} E[Z^-] = E[Z^+] &= E[ZI_{[Z \geq 0]}] \\ &\leq E^{1/4}[Z^4]E^{3/4}[I^{4/3}_{[Z \geq 0]}] = \xi p^{3/4}. \end{aligned}$$

Combining these three inequalities with (9.19) gives $s^2 \leq p^{1/2}\xi^2 + (\xi p^{3/4})^{2/3}\xi^{4/3} = 2p^{1/2}\xi^2$. ■

Chernoff's Theorem†

THEOREM 9.3

Let $X_1, X_2, \ldots$ be independent, identically distributed simple random variables satisfying $E[X_n] < 0$ and $P[X_n > 0] > 0$, let $M(t)$ be their common moment generating function, and put $\rho = \inf_t M(t)$. Then

$$\lim_{n\to\infty} \frac{1}{n} \log P[X_1 + \cdots + X_n \geq 0] = \log \rho. \tag{9.20}$$

Proof. Put $Y_n = X_1 + \cdots + X_n$. Then $E[Y_n] < 0$ and $P[Y_n > 0] \geq P^n[X_1 > 0] > 0$, and so the hypotheses of Theorem 9.1 are satisfied. Define ρ_n and τ_n by $\inf_t M_n(t) = M_n(\tau_n) = \rho_n$, where $M_n(t)$ is the moment generating function of Y_n. Since $M_n(t) = M^n(t)$, it follows that $\rho_n = \rho^n$ and $\tau_n = \tau$, where $M(\tau) = \rho$.

Let Z_n be the analogue for Y_n of the Z described by (9.13). Its moment generating function (see (9.14)) is $M_n(\tau + t)/\rho^n = (M(\tau + t)/\rho)^n$. This is also the

†This theorem is not needed for the law of the iterated logarithm, Theorem 9.5.

moment generating function of $V_1 + \cdots + V_n$ for independent random variables $V_1, \ldots, V_n$ each having moment generating function $M(\tau + t)/\rho$. Now each V_i has (see (9.15)) mean 0 and some positive variance σ^2 and fourth moment ξ^4 independent of i. Since Z_n must have the same moments as $V_1 + \cdots + V_n$, it has mean 0, variance $s_n^2 = n\sigma^2$, and fourth moment $\xi_n^4 = n\xi^4 + 3n(n-1)\sigma^4 = O(n^2)$ (see (6.2)). By Theorem 9.2, $P[Z_n \geq 0] \geq s_n^4/4\xi_n^4 \geq \alpha$ for some positive α independent of n. By Theorem 9.1 then, $P[Y_n \geq 0] = \rho^n e^{-\theta_n}$, where $0 \leq \theta_n \leq \tau_n s_n \alpha^{-1} - \log \alpha = \tau\alpha^{-1}\sigma\sqrt{n} - \log \alpha$. This gives (9.20), and shows, in fact, that the rate of convergence is $O(n^{-1/2})$. ■

This result is important in the theory of statistical hypothesis testing. An informal treatment of the Bernoulli case will illustrate the connection.

Suppose $S_n = X_1 + \cdots + X_n$, where the X_i are independent and assume the values 1 and 0 with probabilities p and q. Now $P[S_n \geq na] = P\left[\sum_{k=1}^n (X_k - a) \geq 0\right]$, and Chernoff's theorem applies if $p < a < 1$. In this case $M(t) = E[e^{t(X_1 - a)}] = e^{-ta}(pe^t + q)$. Minimizing this shows that the ρ of Chernoff's theorem satisfies

$$-\log \rho = K(a,\ p) = a \log \frac{a}{p} + b \log \frac{b}{q},$$

where $b = 1 - a$. By (9.20), $n^{-1} \log P[S_n \geq na] \to -K(a,\ p)$; express this as

$$P[S_n \geq na] \approx e^{-nK(a,\ p)}. \tag{9.21}$$

Suppose now that p is unknown and that there are two competing hypotheses concerning its value, the hypothesis H_1 that $p = p_1$ and the hypothesis H_2 that $p = p_2$, where $p_1 < p_2$. Given the observed results $X_1, \ldots, X_n$ of n Bernoulli trials, one decides in favor of H_2 if $S_n \geq na$ and in favor of H_1 if $S_n < na$, where a is some number satisfying $p_1 < a < p_2$. The problem is to find an advantageous value for the threshold a.

By (9.21),

$$P[S_n \geq na | H_1] \approx e^{-nK(a,\ p_1)}, \tag{9.22}$$

where the notation indicates that the probability is calculated for $p = p_1$—that is, under the assumption of H_1. By symmetry,

$$P[S_n < na | H_2] = e^{-nK(a,\ p_2)}. \tag{9.23}$$

The left sides of (9.22) and (9.23) are the probabilities of erroneously deciding in favor of H_2 when H_1 is, in fact, true and of erroneously deciding in favor of H_1 when H_2 is, in fact, true—the probabilities describing the level and power of the test.

Suppose a is chosen so that $K(a,\ p_1) = K(a,\ p_2)$, which makes the two error probabilities approximately equal. This constraint gives for a a linear equation with solution

$$a = a(p_1, p_2) = \frac{\log(q_1/q_2)}{\log(p_2/p_1) + \log(q_1/q_2)}, \tag{9.24}$$

where $q_i = 1 - p_i$. The common error probability is approximately $e^{-nK(a,\ p_1)}$ for this value of a, and so the larger $K(a,\ p_1)$ is, the easier it is to distinguish statistically between p_1 and p_2.

Although $K(a(p_1, p_2), p_1)$ is a complicated function, it has a simple approximation for p_1 near p_2. As $x \to 0, \log(1 + x) = x - \frac{1}{2}x^2 + O(x^3)$. Using this in the definition of K and collecting terms gives

$$K(p + x, p) = \frac{x^2}{2pq} + O(x^3), \quad x \to 0. \tag{9.25}$$

Fix $p_1 = p$, and let $p_2 = p + t$; (9.24) becomes a function $\psi(t)$ of t, and expanding the logarithms gives

$$\psi(t) = p + \frac{1}{2}t + O(t^2), \quad t \to 0, \tag{9.26}$$

after some reductions. Finally, (9.25) and (9.26) together imply that

$$K(\psi(t), p) = \frac{t^2}{8pq} + O(t^3), \quad t \to 0. \tag{9.27}$$

In distinguishing $p_1 = p$ from $p_2 = p + t$ for small t, if a is chosen to equalize the two error probabilities, then their common value is about $e^{-nt^2/8pq}$. For t fixed, the nearer p is to $\frac{1}{2}$, the larger this probability is and the more difficult it is to distinguish p from $p+t$. As an example, compare $p = .1$ with $p = .5$. Now $.36nt^2/8(.1)(.9) = nt^2/8(.5)(.5)$. With a sample only 36 percent as large, .1 can therefore be distinguished from $.1 + t$ with about the same precision as .5 can be distinguished from $.5 + t$.

The Law of the Iterated Logarithm

The analysis of the rate at which S_n/n approaches the mean depends on the following variant of the theorem on large deviations.

THEOREM 9.4

Let $S_n = X_1 + \cdots + X_n$, where the X_n are independent and identically distributed simple random variables with mean 0 *and variance* 1. *If a_n are constants satisfying*

$$a_n \to \infty, \quad \frac{a_n}{\sqrt{n}} \to 0, \tag{9.28}$$

then

$$P[S_n \geq a_n\sqrt{n}] = e^{-a_n^2(1+\zeta_n)/2} \tag{9.29}$$

for a sequence ζ_n *going to* 0.

Proof. Put $Y_n = S_n - a_n\sqrt{n} = \sum_{k=1}^{n}(X_k - a_n/\sqrt{n})$. Then $E[Y_n] < 0$. Since X_1 has mean 0 and variance 1, $P[X_1 > 0] > 0$, and it follows by (9.28) that $P[X_1 > a_n/\sqrt{n}] > 0$ for n sufficiently large, in which case $P[Y_n > 0] \geq P^n[X_1 - a_n/\sqrt{n} > 0] > 0$. Thus Theorem 9.1 applies to Y_n for all large enough n.

Let $M_n(t), \rho_n, \tau_n$, and Z_n be associated with Y_n as in the theorem. If $m(t)$ and $c(t)$ are the moment and cumulant generating functions of the X_n, then $M_n(t)$ is the nth power of the moment generating function $e^{-ta_n/\sqrt{n}}m(t)$ of $X_1 - a_n/\sqrt{n}$, and so Y_n has cumulant generating function

$$C_n(t) = -ta_n\sqrt{n} + nc(t). \tag{9.30}$$

Since τ_n is the unique minimum of $C_n(t)$, and since $C_n'(t) = -a_n\sqrt{n} + nc'(t), \tau_n$ is determined by the equation $c'(\tau_n) = a_n/\sqrt{n}$. Since X_1 has mean 0 and variance 1, it follows by (9.6) that

$$c(0) = c'(0) = 0. \quad c''(0) = 1. \tag{9.31}$$

Now $c'(t)$ is nondecreasing because $c(t)$ is convex, and since $c'(\tau_n) = a_n/\sqrt{n}$ goes to 0, τ_n must therefore go to 0 as well and must in fact be $O(a_n/\sqrt{n})$. By the second-order mean-value theorem for $c'(t), a_n/\sqrt{n} = c'(\tau_n) = \tau_n + O(\tau_n^2)$, from which follows

$$\tau_n = \frac{a_n}{\sqrt{n}} + O\left(\frac{a_n^2}{n}\right). \tag{9.32}$$

By the third-order mean-value theorem for $c(t)$,

$$\begin{aligned}\log \rho_n &= C_n(\tau_n) = -\tau_n a_n\sqrt{n} + nc(\tau_n) \\ &= -\tau_n a_n\sqrt{n} + n\left[\frac{1}{2}\tau_n^2 + O\left(\tau_n^3\right)\right].\end{aligned}$$

Applying (9.32) gives

$$\log \rho_n = -\frac{1}{2}a_n^2 + o(a_n^2). \tag{9.33}$$

Now (see (9.14)) Z_n has moment generating function $M_n(\tau_n + t)/\rho_n$ and (see (9.30)) cumulant generating function $D_n(t) = C_n(\tau_n + t) - \log \rho_n =$

$-(\tau_n + t)\alpha_n\sqrt{n} + nc(t + \tau_n) - \log \rho_n$. The mean of Z_n is $D_n'(0) = 0$. Its variance s_n^2 is $D_n''(0)$; by (9.31), this is

$$s_n^2 = nc''(\tau_n) = n(c''(0) + O(\tau_n)) = n(1 + o(1)). \tag{9.34}$$

The fourth cumulant of Z_n is $D_n''''(0) = nc''''(\tau_n) = O(n)$. By the formula (9.9) relating moments and cumulants (applicable because $E[Z_n] = 0$), $E[Z_n^4] = 3s_n^4 + D_n''''(0)$. Therefore, $E[Z_n^4]/s_n^4 \to 3$, and it follows by Theorem 9.2 that there exists an α such that $P[Z_n \geq 0] \geq \alpha > 0$ for all sufficiently large n.

By Theorem 9.1, $P[Y_n \geq 0] = \rho_n e^{-\theta_n}$ with $0 \leq \theta_n \leq \tau_n s_n \alpha^{-1} + \log \alpha$. By (9.28), (9.32), and (9.34), $\theta_n = O(a_n) = o(a_n^2)$, and it follows by (9.33) that $P[Y_n \geq 0] = e^{-a_n^2(1+o(1))/2}$. ■

The law of the iterated logarithm is this:

THEOREM 9.5

Let $S_n = X_1 + \cdots + X_n$, where the X_n are independent, identically distributed simple random variables with mean 0 *and variance* 1. *Then*

$$P\left[\limsup_n \frac{S_n}{\sqrt{2n \log\log n}} = 1\right] = 1. \tag{9.35}$$

Equivalent to (9.35) is the assertion that for positive ϵ

$$P[S_n \geq (1 + \epsilon)\sqrt{2n \log\log n} \text{ i.o.}] = 0 \tag{9.36}$$

and

$$P[S_n \geq (1 - \epsilon)\sqrt{2n \log\log n} \text{ i.o.}] = 1. \tag{9.37}$$

The set in (9.35) is, in fact, the intersection over positive rational ϵ of the sets in (9.37) minus the union over positive rational ϵ of the sets in (9.36).

The idea of the proof is this. Write

$$\phi(n) = \sqrt{2n \log\log n}. \tag{9.38}$$

If $A_n^{\pm} = [S_n \geq (1 \pm \epsilon)\phi(n)]$, then by (9.29), $P(A_n^{\pm})$ is near $(\log n)^{-(1\pm\epsilon)^2}$. If n_k increases exponentially, say $n_k \sim \theta^k$ for $\theta > 1$, then $P(A_{n_k}^{\pm})$ is of the order $k^{-(1+\epsilon)^2}$. Now $\sum_k k^{-(1\pm\epsilon)^2}$ converges if the sign is $+$ and diverges if the sign is $-$. It will follow by the first Borel–Cantelli lemma that there is probability 0 that $A_{n_k}^+$ occurs for infinitely many k. In providing (9.36), an extra argument is required to get around the fact that the A_n^+ for $n \neq n_k$ must also be accounted for (this requires choosing θ near 1). If the A_n^- were independent, it would

follow by the second Borel–Cantelli lemma that with probability 1, $A_{n_k}^-$ occurs for infinitely many k, which would in turn imply (9.37). An extra argument is required to get around the fact that the $A_{n_k}^-$ are dependent (this requires choosing θ large).

For the proof of (9.36) a preliminary result is needed. Put $M_k = \max\{S_0, S_1, \ldots, S_k\}$, where $S_0 = 0$.

THEOREM 9.6

If the X_k are independent simple random variables with mean 0 and variance 1, then for $\alpha \geq \sqrt{2}$.

$$P\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right] \leq 2P\left[\frac{S_n}{\sqrt{n}} \geq \alpha - \sqrt{2}\right]. \tag{9.39}$$

Proof. If $A_j = [M_{j-1} < \alpha\sqrt{n} \leq M_j]$, then

$$P\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right] \leq P\left[\frac{S_n}{\sqrt{n}} \geq \alpha - \sqrt{2}\right] + \sum_{j=1}^{n-1} P\left(A_j \cap \left[\frac{S_n}{\sqrt{n}} \leq \alpha - \sqrt{2}\right]\right).$$

Since $S_n - S_j$ has variance $n-j$, it follows by independence and Chebyshev's inequality that the probability in the sum is at most

$$\begin{aligned} P\left(A_j \cap \left[\frac{|S_n - S_j|}{\sqrt{n}} > \sqrt{2}\right]\right) &= P(A_j)P\left[\frac{|S_n - S_j|}{\sqrt{n}} > \sqrt{2}\right] \\ &\leq P(A_j)\frac{n-j}{2n} \leq \frac{1}{2}P(A_j). \end{aligned}$$

Since $\bigcup_{j=1}^{n-1} A_j \subset [M_n \geq \alpha\sqrt{n}]$,

$$P\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right] \leq P\left[\frac{S_n}{\sqrt{n}} \geq \alpha - \sqrt{2}\right] + \frac{1}{2}P\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right].$$

■

Proof of (9.36). Given ϵ, choose θ so that $\theta > 1$ but $\theta^2 < 1 + \epsilon$. Let $n_k = \lfloor \theta^k \rfloor$ and $x_k = \theta(2\log\log n_k)^{1/2}$. By (9.29) and (9.39),

$$P\left[\frac{M_{n_k}}{\sqrt{n_k}} \geq x_k\right] \leq 2\exp\left[-\frac{1}{2}(x_k - \sqrt{2})^2(1 + \xi_k)\right].$$

where $\xi_k \to 0$. The negative of the exponent is asymptotically $\theta^2 \log k$ and hence for large k exceeds $\theta \log k$, so that

$$P\left[\frac{M_{n_k}}{\sqrt{n_k}} \geq x_k\right] \leq \frac{2}{k^\theta}.$$

Since $\theta > 1$, it follows by the first Borel–Cantelli lemma that there is probability 0 that (see (9.38))

$$M_{n_k} \geq \theta\phi(n_k) \tag{9.40}$$

for infinitely many k. Suppose that $n_{k-1} < n \leq n_k$ and that

$$S_n > (1+\epsilon)\phi(n). \tag{9.41}$$

Now $\phi(n) \geq \phi(n_{k-1}) \sim \theta^{-1/2}\phi(n_k)$; hence, by the choice of θ, $(1+\epsilon)\phi(n) > \theta\phi(n_k)$ if k is large enough. Thus for sufficiently large k, (9.41) implies (9.40) (if $n_{k-1} < n \leq n_k$), and there is therefore proability 0 that (9.41) holds for infinitely many n. ■

Proof of (9.37). Given ϵ, choose an integer θ so large that $3\theta^{-1/2} < \epsilon$. Take $n_k = \theta^k$. Now $n_k - n_{k-1} \to \infty$, and (9.29) applies with $n = n_k - n_{k-1}$ and $a_n = x_k/\sqrt{n_k - n_{k-1}}$, where $x_k = (1-\theta^{-1})\phi(n_k)$. It follows that

$$P[S_{n_k} - S_{n_{k-1}} \geq x_k] = P[S_{n_k - n_{k-1}} \geq x_k] = \exp\left[-\frac{1}{2}\frac{x_k^2}{n_k - n_{k-1}}(1+\xi_k)\right],$$

where $\xi_k \to 0$. The negative of the exponent is asymptotically $(1-\theta^{-1})\log k$ and so for large k is less than $\log k$, in which case $P[S_{n_k} - S_{n_{k-1}} \geq x_k] \geq k^{-1}$. The events here being independent, it follows by the second Borel–Cantelli lemma that with probability 1, $S_{n_k} - S_{n_{k-1}} \geq x_k$ for infinitely many k. On the other hand, by (9.36) applied to $\{-X_n\}$, there is probability 1 that $-S_{n_{k-1}} \leq 2\phi(n_{k-1}) \leq 2\theta^{-1/2}\phi(n_k)$ for all but finitely many k. These two inequalities give $S_{n_k} \geq x_k - 2\theta^{-1/2}\phi(n_k) > (1-\epsilon)\phi(n_k)$, the last inequality because of the choice of θ. ■

That completes the proof of Theorem 9.5.

PROBLEMS

9.1. Prove (6.2) by using (9.9) and the fact that cumulants add in the presence of independence.

9.2. In the Bernoulli case, (9.21) gives

$$P[S_n \geq np + x_n] = \exp\left[-nK\left(p + \frac{x_n}{n}, p\right)(1+o(1))\right],$$

where $p < a < 1$ and $x_n = n(a - p)$. Theorem 9.4 gives

$$P[S_n \geq np + x_n] = \exp\left[-\frac{x_n^2}{2npq}(1 + o(1))\right),$$

where $x_n = a_n\sqrt{npq}$. Resolve the apparent discrepancy. Use (9.25) to compare the two expressions in case x_n/n is small. See Problem 27.17.

9.3. Relabel the binomial parameter p as $\theta = f(p)$, where f is increasing and continuously differentiable. Show by (9.27) that the distinguishability of θ from $\theta + \Delta\theta$, as measured by K, is $(\Delta\theta)^2/8p(1 - p)(f'(p))^2 + O(\Delta\theta)^3$. The leading coefficient is independent of θ if $f(p) = \arcsin\sqrt{p}$.

9.4. From (9.35) and the same result for $\{-X_n\}$, together with the uniform boundedness of the X_n, deduce that with probability 1 the set of limit points of the sequence $\{S_n(2n \log\log n)^{-1/2}\}$ is the closed interval from -1 to $+1$.

9.5. ↑ Suppose X_n takes the values ± 1 with probability $\frac{1}{2}$ each, and show that $P[S_n = 0 \text{ i.o.}] = 1$. (This gives still another proof of the persistence of symmetric random walk on the line (Example 8.6).) Show more generally that, if the X_n are bounded by M, then $P[|S_n| \leq M \text{ i.o.}] = 1$.

9.6. Weakened versions of (9.36) are quite easy to prove. By a fourth-moment argument (see (6.2)), show that $P[S_n > n^{3/4}(\log n)^{(1+\epsilon)/4} \text{ i.o.}] = 0$. Use (9.29) to give a simple proof that $P[S_n > (3n \log n)^{1/2} \text{ i.o.}] = 0$.

9.7. Show that (9.35) is true if S_n is replaced by $|S_n|$ or $\max_{k \leq n} S_k$ or $\max_{k \leq n} |S_k|$.