We will not prove Theorem 5.1 in its full generality as it requires extensive measure theoretic considerations. We refer the reader to Doob [1953, Chapter 7], or Neveu [1965]. The upcrossings inequality of the last section is the key tool. We will secure, however, a convergence theorem whose proof parallels and illustrates the general approach but is somewhat simpler in conception. Specifically, the convergence Theorem 5.2 concerns martingales satisfying the stronger condition (5.6) for $\rho = 1$. Concomitantly, under this strengthened hypothesis, is the sharper conclusion to the effect that besides convergence of X_n to X_∞ with probability one, convergence in mean square also prevails. The maximal inequality discussed next is a basic tool for this end and also serves in numerous other capacities.

THE MAXIMAL INEQUALITY

Let $\xi_1, \xi_2, \xi_3, \ldots$ be independent and identically distributed random variables obeying the moment conditions $E[\xi_i] = 0$ and $E[\xi_i^2] = \sigma^2 < \infty$. Define

$$S_0 = 0, \quad \text{and} \quad S_n = \xi_1 + \cdots + \xi_n,$$

for $n \geq 1$. Noting that the variance of S_n is $n\sigma^2$, Chebyshev's inequality gives

$$\varepsilon^2 \Pr\{|S_n| > \varepsilon\} \leq n\sigma^2, \qquad \varepsilon > 0.$$

A finer inequality is available:

$$\varepsilon^2 \Pr\left\{\max_{0 \leq k \leq n} |S_k| > \varepsilon\right\} \leq n\sigma^2, \tag{5.7}$$

known as *Kolmogorov's inequality*. The generalization to submartingales is the simple, yet powerful, *maximal inequality for submartingales*. (See Problem 5 for a strengthened version.)

Lemma 5.1. *Let $\{X_n\}$ be a submartingale for which $X_n \geq 0$ for all n. Then for any positive λ.*

$$\lambda \Pr\left\{\max_{0 \leq k \leq n} X_k > \lambda\right\} \leq E[X_n]. \tag{5.8}$$

Proof. Define the Markov time

$$T = \begin{cases} \min\{k \geq 0;\ X_k > \lambda\}, & \text{if } X_k > \lambda, \text{ for some } k = 0, \ldots, n, \\ n, & \text{if } X_k \leq \lambda, \text{ for } k = 0, 1, \ldots, n, \end{cases}$$

Applying the submartingale analog of Lemma 3.2, the optional stopping lemma, yields

$$\begin{aligned} E[X_n] &\geq E[X_T] \\ &\geq E\left[X_T \cdot I\left\{\max_{0\leq k\leq n} X_k > \lambda\right\}\right] \qquad \text{(since } X_i \text{ are all nonnegative)} \\ &\geq \lambda \Pr\left\{\max_{0\leq k\leq n} X_k > \lambda\right\} \qquad \text{(since on the indicator set, } X_T \geq \lambda\text{)}, \end{aligned}$$

the desired inequality. ■

Kolmogorov's inequality ensues immediately, since $\{S_n\}$ is a martingale [Example (a) of Section 1], and, according to Lemma 2.1, $X_n = S_n^2$ determines a submartingale, obviously nonnegative, whence

$$\begin{aligned} n\sigma^2 = E[S_n^2] \\ &\geq \lambda \Pr\left\{\max_{0\leq k\leq n} S_k^2 > \lambda\right\} \\ &= \varepsilon^2 \Pr\left\{\max_{0\leq k\leq n} |S_k| > \varepsilon\right\}, \qquad \text{for} \quad \varepsilon = \sqrt{\lambda}. \end{aligned}$$

Corollary 5.1. *Let $\{X_n\}$ be a martingale. Then for every positive λ*

$$\lambda \Pr\left\{\max_{0\leq k\leq n} |X_k| > \lambda\right\} \leq E[|X_n|].$$

Proof. If $\{X_n\}$ is a martingale, then Lemma 2.1 assures us that $\{|X_n|\}$ is a nonnegative submartingale. The maximal inequality just proved then applies. ■

The proof of Lemma 5.1 may be readily adapted to yield the *maximal inequality for supermartingales*, whose statement follows. (See Problem 12.)

Lemma 5.2. *If $\{X_n\}$ is a nonnegative supermartingale, then*

$$\lambda \Pr\left\{\max_{0\leq k\leq n} X_k > \lambda\right\} \leq E[X_0], \qquad \text{for} \quad \lambda > 0.$$

Example. Define $X_0 = 1$ and $X_n = \prod_{i=1}^{n} Y_i$, for $n \geq 1$, where $Y_1, Y_2, \ldots$ are nonnegative independent random variables having a common unit mean, $E[Y_i] = 1$. Then $\{X_n\}$ is a nonnegative martingale, and the maximal inequality says

$$\Pr\left\{\max_{0\leq k\leq n} X_k > \lambda\right\} \leq 1/\lambda, \qquad \text{for} \quad \lambda > 0. \tag{*}$$

This bound is rather frustrating as the following situation will illustrate. Consider a gambler who risks a fraction q of his fortune, $0 < q < 1$, with each toss of a fair coin. Starting with one dollar, straightforward induction verifies that his fortune X_n after n tosses is $\prod_{j=1}^{n} (1 + \delta_j q)$, where δ_j is a sequence of independent random variables with possible values ± 1, each occurring with probability $\frac{1}{2}$. Now, if our gambler is patient enough, he will see his fortune dwindle to zero. Since $\{X_n\}$ is a martingale, it is a fortiori a submartingale, and, being positive,

$$\sup_{n \geq 1} E[|X_n|] = E[X_n] = 1,$$

so that (5.1) is satisfied. Thus with probability one, X_n tends to a finite limit as $n \to \infty$, which must be zero, since every other state is transient. The inequality of (*) appears rather weak in this context.

THE MARTINGALE MEAN SQUARE CONVERGENCE THEOREM

Let $\{X_n\}$ be a martingale and let *A be the random event that the sequence $\{X_n\}$ converges*. A formal characterization of the set A will be forthcoming. A particular realization $X_0, X_1, \ldots$ converges if and only if the Cauchy criterion

$$\lim_{m, n \to \infty} |X_m - X_n| = 0$$

is satisfied, and thus A has the explicit form

$$A = \left\{ \lim_{m, n \to \infty} |X_m - X_n| = 0 \right\}. \tag{5.9}$$

In words, A is the event that the process realization $X_0, X_1, \ldots$ satisfies the Cauchy criterion for convergence. When A occurs, let X_∞ denote the limit. We want to show $\Pr\{A\} = 1$. Then X_∞ will be defined, not always, but at least for a set of realizations $X_0, X_1, \ldots$ having total probability one, and

$$\Pr\left\{ \lim_{n \to \infty} X_n = X_\infty \right\} = 1.$$

Under the assumption that the second moments of $\{X_n\}$ are uniformly bounded, we will indeed prove convergence with probability one and also that convergence in mean square takes place.

Theorem 5.2. *Let $\{X_n\}$ be a martingale with respect to $\{Y_n\}$ satisfying, for some constant K,*

$$E[X_n^2] \leq K < \infty, \qquad \textit{for all} \quad n. \tag{5.10}$$

Then $\{X_n\}$ *converges as* $n \to \infty$ *to a limit random variable* X_∞ *both with probability one and in mean square. That is,*

$$\Pr\left\{\lim_{n\to\infty} X_n = X_\infty\right\} = 1, \tag{5.11}$$

and

$$\lim_{n\to\infty} E[|X_n - X_\infty|^2] = 0, \tag{5.12}$$

prevail. Finally,

$$E[X_0] = E[X_n] = E[X_\infty], \qquad \text{for all} \quad n. \tag{5.13}$$

Proof. Temporarily fix N, and for $k \geq 0$ set

$$\tilde{X}_k = X_{N+k} - X_N.$$

Now, the law of total probability and appropriate conditioning gives

$$\begin{aligned} E[X_{N+k}X_N] &= E\{E[X_{N+k}X_N|Y_0, \ldots, Y_N]\} \\ &= E\{X_N E[X_{N+k}|Y_0, \ldots, Y_N]\} \\ &= E[X_N^2], \end{aligned}$$

so that

$$\begin{aligned} 0 < E[\tilde{X}_k^2] &= E[(X_{N+k} - X_N)^2] \\ &= E[X_{N+k}^2 - 2X_{N+k}X_N + X_N^2] \\ &= E[X_{N+k}^2] - E[X_N^2]. \end{aligned} \tag{5.14}$$

Lemma 2.1 tells us that $\{X_n^2\}$ is a submartingale, $\{X_n\}$ being a martingale, and (5.14) indicates that $E[X_n^2]$ is a monotone nondecreasing sequence, bounded above by K, and hence convergent. Accordingly, the Cauchy criterion applies to give

$$\begin{aligned} 0 &= \lim_{N,k\to\infty} \{E[X_{N+k}^2] - E[X_N^2]\} \\ &= \lim_{N,k\to\infty} E[\tilde{X}_k^2]. \end{aligned} \tag{5.15}$$

We will use this in a minute.

Let A be the event that $\{X_n\}$ converges. Explicitly,

$$\begin{aligned} A = {} & \text{set of all realizations where } \lim_{m,n\to\infty} |X_m - X_n| = 0 \\ = {} & \text{set of realizations } \{X_n\} \text{ for which, for every } \varepsilon > 0, \text{ there} \\ & \text{exists } N > 0 \text{ satisfying } |X_{N+m} - X_{N+n}| \leq \varepsilon \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{for all} \quad m, n \geq 1. \end{aligned}$$

From the triangle inequality

$$|X_{N+m} - X_{N+n}| \leq |X_{N+m} - X_N| + |X_{N+n} - X_N|,$$

we see that A may be described equivalently in the terms

$$A = \{\text{For every } \varepsilon > 0 \text{ there exists } N \geq 0 \text{ for which } |X_{N+k} - X_N| \leq \varepsilon \text{ for all } k \geq 0\}.$$

Let B denote the complementary event to A consisting of the realizations where $\{X_n\}$ does not converge. Then

$$\begin{aligned} B = &\{\text{for some } \varepsilon > 0, \text{ for every } N \geq 0, \\ &\text{there exists } k \geq 0 \text{ depending on } \varepsilon, N \text{ and the realization} \\ &\text{for which } |X_{N+k} - X_N| > \varepsilon\} \\ = &\bigcup_{\varepsilon > 0} \{\text{for every } N \geq 0 \text{ there exists } k \geq 0 \\ &\text{for which } |X_{N+k} - X_N| > \varepsilon\} \\ = &\bigcup_{\varepsilon > 0} \bigcap_{N=0}^{\infty} B_N(\varepsilon), \end{aligned}$$

where

$$B_N(\varepsilon) = \text{event described by the conditions } \{|X_{N+k} - X_N| > \varepsilon \text{ for some } k = 0, 1, \ldots\}.$$

We wish to prove the equation $\Pr\{B\} = 0$. For this objective it suffices to establish

$$\lim_{N \to \infty} \Pr\{B_N(\varepsilon)\} = 0, \qquad \text{for every} \quad \varepsilon > 0, \tag{5.16}$$

since in that case

$$\begin{aligned} \Pr\{B\} &= \Pr\left\{ \bigcup_{\varepsilon > 0} \bigcap_{N=0}^{\infty} B_N(\varepsilon) \right\} \\ &= \lim_{\varepsilon \downarrow 0} \Pr\left\{ \bigcap_{N=0}^{\infty} B_N(\varepsilon) \right\} \\ &= \lim_{\varepsilon \downarrow 0} \lim_{N \to \infty} \Pr\{B_N(\varepsilon)\} = 0. \end{aligned}$$

To validate (5.16) we will employ the maximal inequality. Fix N and put

$$\tilde{X}_k = X_{N+k} - X_N, \qquad k = 0, 1, \ldots,$$
$$\tilde{Y}_0 = (Y_0, \ldots, Y_N),$$

and

$$\tilde{Y}_k = Y_{N+k}, \qquad k = 1, 2, \ldots.$$

From Jensen's inequality, or more specifically Schwarz' inequality, and with the stipulation of (5.14), we obtain

$$E[|\tilde{X}_k|] \leq (E[\tilde{X}_k^2])^{1/2} \leq \sqrt{K} < \infty,$$

and

$$\begin{aligned} E[\tilde{X}_{k+1}|\tilde{Y}_0, \ldots, \tilde{Y}_k] &= E[X_{N+k+1} - X_N|Y_0, \ldots, Y_{N+k}] \\ &= X_{N+k} - E[X_N|Y_0, \ldots, Y_{N+k}] \\ &= X_{N+k} - X_N = \tilde{X}_k, \end{aligned}$$

so that $\{\tilde{X}_k\}$ is a martingale with respect to $\{\tilde{Y}_k\}$. Again, Lemma 2.1 tells us that $\{\tilde{X}_k^2\}$ is a submartingale, and the maximal inequality yields

$$\varepsilon^2 \Pr\left\{\max_{0 \leq k \leq n} |\tilde{X}_k^2| > \varepsilon^2\right\} \leq E[\tilde{X}_n^2],$$

or

$$\varepsilon^2 \Pr\left\{\max_{0 \leq k \leq n} |X_{N+k} - X_N| > \varepsilon\right\} \leq E[\tilde{X}_n^2].$$

But in (5.15) we showed that the right-hand side goes to zero as $N, n \to \infty$. It follows that

$$\begin{aligned} 0 &= \lim_{N \to \infty} \lim_{n \to \infty} \Pr\left\{\max_{0 \leq k \leq n} |X_{N+k} - X_N| > \varepsilon\right\} \\ &= \lim_{N \to \infty} \Pr\left\{\sup_{0 \leq k < \infty} |X_{N+k} - X_N| > \varepsilon\right\} \\ &= \lim_{N \to \infty} \Pr\{B_N(\varepsilon)\}. \end{aligned}$$

The preceding considerations complete the proof for the convergence of $\{X_n\}$ with probability one.

Let X_∞ denote the limit random variable. It remains to verify that $\{X_n\}$ converges to X_∞ in mean square. This requires only justification of the second inequality in

$$\begin{aligned} 0 &\leq \lim_{n \to \infty} E[|X_n - X_\infty|^2] \\ &= \lim_{n \to \infty} E[\lim_{m \to \infty} |X_n - X_m|^2] \\ &\leq \lim_{n \to \infty} \lim_{m \to \infty} E[|X_n - X_m|^2] = 0. \end{aligned}$$

The last limit is zero, of course, owing to (5.15). Fix n and let $Z_m = |X_n - X_m|^2$, so that we want to prove

$$\lim_{m \to \infty} E[Z_m] \geq E\left[\lim_{m \to \infty} Z_m\right]. \tag{5.17}$$

We begin with the representation

$$\begin{aligned} E[Z_m] &= \int_0^\infty \Pr\{Z_m \geq t\}\, dt \\ &= \sum_{k=1}^{\infty} \int_{(k-1)\varepsilon}^{k\varepsilon} \Pr\{Z_m \geq t\}\, dt \\ &\geq \sum_{k=1}^{\infty} \int_{(k-1)\varepsilon}^{k\varepsilon} \Pr\{Z_m \geq k\varepsilon\}\, dt \\ &\geq \sum_{k=1}^{N} \varepsilon \Pr\{Z_m \geq k\varepsilon\}, \end{aligned} \tag{5.18}$$

where $\varepsilon > 0$, $N > 0$ are arbitrary.

Remark 5.2 recalled the property that convergence with probability one entails convergence in probability, whence, for $\delta > 0$,

$$\begin{aligned} \lim_{m\to\infty} \Pr\{Z_m \geq k\varepsilon\} &\geq \lim_{m\to\infty} [\Pr\{Z \geq k\varepsilon + \delta\} - \Pr\{|Z_m - Z| > \delta\}] \\ &= \Pr\{Z \geq k\varepsilon + \delta\}, \end{aligned}$$

where $Z = \lim_{m\to\infty} Z_m$.

Since $\delta > 0$ is arbitrary,

$$\lim_{m\to\infty} \Pr\{Z_m \geq k\varepsilon\} \geq \Pr\{Z > k\varepsilon\}.$$

Now returning to (5.18), we have

$$\begin{aligned} \lim_{m\to\infty} E[Z_m] &\geq \lim_{m\to\infty} \sum_{k=1}^{N} \varepsilon \Pr\{Z_m \geq k\varepsilon\} \\ &\geq \sum_{k=1}^{N} \varepsilon \Pr\{Z > k\varepsilon\} \\ &\geq \sum_{k=0}^{N} \varepsilon \Pr\{Z > k\varepsilon\} - \varepsilon \\ &\geq \sum_{k=0}^{N} \int_{k\varepsilon}^{(k+1)\varepsilon} \Pr\{Z > t\}\, dt - \varepsilon \\ &\geq \int_0^{(N+1)\varepsilon} \Pr\{Z > t\}\, dt - \varepsilon. \end{aligned}$$

Keep $\varepsilon > 0$ fixed, and let $N \to \infty$ to deduce

$$\lim_{m\to\infty} E[Z_m] \geq \int_0^\infty \Pr\{Z > t\}\, dt - \varepsilon = E[Z] - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, (5.17) is verified, and $\{X_n\}$ converges to X_∞ in mean square. This implies convergence in the mean, by Schwarz' inequality, viz.,

$$0 \leq E[|X_n - X_\infty|] \leq \{E[|X_n - X_\infty|^2]\}^{1/2} \to 0, \qquad \text{as} \quad n \to \infty.$$

And, in the same vein, we obtain

$$\begin{aligned} 0 &\leqq |E[X_n] - E[X_\infty]| \\ &= |E[X_n - X_\infty]| \leq E[|X_n - X_\infty|] \to 0, \qquad \text{as} \quad n \to \infty, \end{aligned}$$

so that $E[X_n] \to E[X_\infty]$ But $E[X_n] = E[X_0]$, for all n. Therefore,

$$E[X_0] = E[X_n] = E[X_\infty]$$

This completes the proof of the martingale mean square convergence theorem ■

6: *Applications and Extensions of the Martingale Convergence Theorems*

Here are some sample implications of the convergence theorems of Section 5.

(a) *Bounded Solutions of $y = Py$, where $P = \|P_{ij}\|$ is the Transition Matrix of an Irreducible Recurrent Markov Chain $\{Y_n\}$.* The martingale convergence theorem can be invoked to establish that every bounded solution $y = \{y(i)\}$ to

$$y(i) = \sum_{j=0}^{\infty} P_{ij}\, y(j), \qquad \text{for all} \quad i, \quad \text{is constant,}$$

i.e., $y(i) = y(j)$ for all i, j. (Compare with Chapter 3, Theorem 4.1.) Because $X_n = y(Y_n)$ is a bounded martingale [cf. Example (d) of Section 1], we have $\lim_{n\to\infty} X_n = \lim_{n\to\infty} y(Y_n)$ exists with probability one. Since the chain is recurrent, all states are visited infinitely often (see Chapter 2), and so

$$\{X_n = y(i)\}, \qquad \text{and} \qquad \{X_n = y(j)\},$$

necessarily both occur for infinitely many n. However, $\lim_{n\to\infty} X_n$ exists, so we must have $y(i) = y(j)$.

(b) *Solutions to $f(y) = \int f(y+z)p(z)\,dz$.* Let $p(z)$ be a continuous probability density function. Every constant function $f(y) \equiv a$ is a solution to the integral equation

$$f(y) = \int f(y+z)p(z)\,dz, \qquad \text{for all} \quad y. \tag{6.1}$$

The martingale convergence theorem can be used to show that, in fact, the only *bounded and continuous solutions to* (6.1) are constant functions. To see this, suppose $f(y)$ is bounded, continuous, and solves (6.1). Then, for every x, $\{f(x+S_n)\}$ constitutes a martingale sequence, where $\{S_n\}$ designates the sequence of partial sums generated by the independent identically distributed random variables $X_1, X_2, \ldots$, whose common probability density function is p. Since $f(y)$ is bounded, the conditions for the mean square convergence theorem are satisfied, and for each x we infer the existence of a random variable U_x satisfying

$$\Pr\left\{\lim_{n\to\infty} f(x+S_n) = U_x\right\} = 1,$$

and

$$\lim_{n\to\infty} E[|f(x+S_n) - U_x|^2] = 0.$$

We prove first that U_x is not random, but a constant $U_x = u(x)$, and subsequently we will prove that $u(x)$ is actually independent of x.

It is a trivial fact that $f(x+S_m-S_n)$ has the same distribution as $f(x+S_{m-n})$ for $m \geq n$. It is a more recondite property that the pair of random variables

$$\{f(x+S_m), \quad f(x+S_m-S_n)\},$$

shares the identical joint distribution as the pair

$$\{f(x+S_m), \quad f(x+S_{m-n})\}, \quad m \geq n.$$

(Why?) Using these facts, we have

$$\lim_{n\to\infty}\lim_{m\to\infty} E[|f(x+S_m) - f(x+S_m-S_n)|^2] \tag{6.2}$$
$$= \lim_{n\to\infty}\lim_{m\to\infty} E[|f(x+S_m) - f(x+S_{m-n})|^2] = 0.$$

This is the crucial step, as will be seen imminently.

On the basis of Schwarz' inequality, we have

$$\{E[U_x^2 - f(x+S_n)U_x]\}^2 = \{E[U_x\{U_x - f(x+S_n)\}]\}^2$$
$$\leq E[U_x^2] \cdot E[|U_x - f(x+S_n)|^2].$$

But the right-hand side goes to zero as n increases, implying

$$E[U_x^2] = \lim_{n\to\infty} E[f(x+S_n)U_x]$$

Analogously, we deduce

$$E[f(x+S_n)U_x] = \lim_{m\to\infty} E[f(x+S_n)f(x+S_m)].$$

Putting these relations together,

$$E[U_x^2] = \lim_{n\to\infty}\lim_{m\to\infty} E[f(x+S_n)f(x+S_m)]$$
$$= \lim_{n\to\infty}\lim_{m\to\infty} E[f(x+S_n)f(x+S_m-S_n)]$$
$$+ \lim_{n\to\infty}\lim_{m\to\infty} E[f(x+S_n)\{f(x+S_m)-f(x+S_m-S_n)\}].$$

For the first term on the right, since S_n and $S_m - S_n$ are independent and $E[f(x+S_n)] = E[U_x]$ for all n, by the martingale convergence theorem, we have

$$E[f(x+S_n)f(x+S_m-S_n)] = E[f(x+S_n)]E[f(x+S_m-S_n)] = \{E[U_x]\}^2.$$

Examining the second term, we employ Schwarz' inequality to obtain

$$\{E[f(x+S_n)\{f(x+S_m)-f(x+S_m-S_n)\}]\}^2$$
$$\leq E[\{f(x+S_n)\}^2]\cdot E[|f(x+S_m)-f(x+S_m-S_n)|^2],$$

and since $f(y)$ is bounded, the last factor goes to zero as $m\to\infty$, utilizing the crucial observation (6.2).

Combining all these relations leads to the equation

$$E[U_x^2] = \{E[U_x]\}^2,$$

which says that the variance of U_x is zero, and this means that U_x is a nonrandom constant, say $u(x)$.

A martingale has a constant mean, and, because the martingale mean square convergence theorem applies, this mean value is maintained for the limit random variable. Accordingly,

$$f(x) = E[f(x+S_0)] = E[U_x] = u(x).$$

At this stage we have established that, with probability one,

$$\lim_{n\to\infty} f(x+S_n) = u(x) = f(x).$$

But we also necessarily have

$$\lim_{n\to\infty} f(x+X_1+S_n-X_1) = f(x+X_1),$$

so that

$$\Pr\{f(x) = f(x+X_1)\} = 1,$$

and, by induction,

$$\Pr\{f(x) = f(x+S_n)\} = 1.$$

This fact in conjunction with the assumption that X_1 is a continuous random variable having probability density function $p(x)$ may be exploited to prove the identity $f(x) = f(y)$ for all x, y. For example, the desired inference is immediate if $p(z) > 0$ for all z, since

$$\begin{aligned} 0 &= E[|f(x + X_1) - f(x)|] \\ &= \int |f(x+z) - f(x)| p(z)\, dz \end{aligned}$$

requires $|f(x+z) - f(x)| = 0$ for all z, as $f(x)$ is continuous, by hypothesis.

It follows that $f(x) = a$ for all x, and every bounded continuous solution to (6.1) is constant.

If $f(x)$ is assumed to achieve its maximum at a point x_0, a much simpler proof is possible. With this added assumption, (6.1) gives

$$\begin{aligned} 0 &= \int \{f(x_0) - f(x_0 + y)\} p(y)\, dy \\ &= \int |f(x_0) - f(x_0 + y)| p(y)\, dy, \end{aligned}$$

which easily implies $f(x_0) = f(x_0 + y)$ for all y, when $p(y) > 0$ for all y. In contrast, such a simple proof is not possible in the general case.

(c) *An Urn Model.* Consider the urn scheme of Example (i), Section 1, where the urn contains n red balls and m green balls at stage 0. As before, X_k, the fraction of red balls in the urn at stage $k = 0, 1, \ldots$, determines a bounded martingale, so that $X_\infty = \lim_{k \to \infty} X_k$ exists with probability one.

The limit random variable X_∞ has a Beta distribution. We will sketch the derivation. Let $Y_k = (n + m + k)X_k$ be the total number of red balls in the urn at stage k. A straightforward induction on k validates the formula

$$\Pr\{Y_k = i\} = \frac{\binom{i-1}{n-1}\binom{N-i-1}{m-1}}{\binom{N-1}{n+m-1}}, \qquad n \le i \le n + k, \qquad N = n + m + k.$$

Use Stirling's approximation $M! \sim e^{-M} M^M (2\pi M)^{1/2}$ to show that $\binom{M}{j} \sim (M - j)^j / j!$ for M large and j fixed. Then

$$\begin{aligned} \Pr\{X_k \le x\} &= \Pr\{Y_k \le Nx\} \\ &= \sum_{i=0}^{[Nx]} \binom{i-1}{n-1}\binom{N-i-1}{m-1} \Big/ \binom{N-1}{n+m-1} \\ &\cong \frac{(n+m-1)!}{(n-1)!\,(m-1)!} \sum_{i=0}^{[Nx]} \left(\frac{i-n}{k}\right)^{n-1} \left(\frac{k-i+n}{k}\right)^{m-1} \frac{1}{k} \\ &= \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} \sum_{i/k=0}^{[Nx]/k} \left(\frac{i}{k} - \frac{n}{k}\right)^{n-1} \left(1 - \frac{i}{k} + \frac{n}{k}\right)^{m-1} \Delta\left(\frac{i}{k}\right), \end{aligned}$$

where $\Delta(i/k) = (i+1)/k - i/k = 1/k$. When N and k are large, n/k becomes negligible, and the sum is approximated by the integral in which $z = i/k$, $dz \approx \Delta(i/k)$, and $[Nx]/k \sim x$. Thus, as $k \to \infty$,

$$\Pr\{X_k \leq x\} \to \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} \int_0^x z^{n-1}(1-z)^{m-1}\, dz$$
$$= \Pr\{X_\infty \leq x\}, \qquad \text{for} \quad 0 \leq x \leq 1.$$

This is the Beta distribution, as asserted.

(d) *Branching Processes.* Example (f) of Section 1 indicated that $X_n = m^{-n}Y_n$ is a martingale, where $m < \infty$ is the mean of the offspring distribution of a branching process $\{Y_n\}$. The basic martingale convergence theorem, amplified by Remark 5.1, tells us that $X_\infty = \lim_{n\to\infty} X_n$ exists with probability one. Roughly speaking, Y_n behaves asymptotically like $X_\infty m^n$, and it is remarkable that the entire asymptotic behavior is captured in the single random variable X_∞.

In Chapter 8 we will show that whenever $m \leq 1$, $Y_n \to 0$ with probability one. Since the possible values for Y_n are $\{0, 1, \ldots\}$, this means in almost every realization of the process, $Y_n = 0$ for sufficiently large n, and therefore

$$X_\infty = \lim_{n\to\infty} m^{-n}Y_n = 0.$$

Thus, here is a case in which

$$1 = E[m^{-n}Y_n] \neq E[X_\infty] = 0,$$

so that the sequence $X_n = m^{-n}Y_n$ cannot be uniformly integrable when $m \leq 1$.

When $m > 1$, and the progeny distribution has a finite variance σ^2, from Section 2 of Chapter 8 we obtain the variance

$$\text{Var}[Y_n] = \frac{\sigma^2}{m}\left(\frac{m^{2n} - m^n}{m-1}\right)$$

and use this to get

$$E[X_n^2] = \{\text{Var}[Y_n] + (E[Y_n])^2\}/m^{2n}$$
$$= \frac{\sigma^2}{m}\left(\frac{1-m^{-n}}{m-1}\right) + 1$$
$$\leq \frac{\sigma^2}{m(m-1)} + 1 \qquad \text{for all } n.$$

We see that the conditions for the martingale mean square convergence theorem are satisfied. Thus $1 = X_0 = E[X_\infty]$, and with positive probability X_∞ is strictly positive, and then $Y_n \sim X_\infty m^n$ which shows that Y_n asymptotically grows exponentially fast at rate m.

(e) *Split Times in Branching Processes.* Consider a population of particles having independent random lifetimes, at the end of which each particle splits into a random number of new particles that independently exhibit the same life behavior as the parent. Suppose specifically that the lifetimes are all independent exponentially distributed random variables with *the same parameter a.*

Let $X(t)$ be the number of particles in the population at time t. Let τ_n be the time of the nth split in the population ($\tau_0 = 0$). The random variable $\xi_n = X(\tau_n + 0) - X(\tau_n - 0)$ counts the number of progeny contributed at the nth split. Let

$$X(0) = 1$$
$$S_i = X(\tau_i) = X(\tau_i + 0) = \xi_1 + \cdots + \xi_i + 1,$$

and define $T_i = \tau_i - \tau_{i-1}$ as the time between the $(i-1)$th and ith splits. We claim that the sequence of random variables

$$Y_n = \sum_{i=1}^{n} \left(T_i - \frac{1}{aS_{i-1}} \right), \qquad n = 1, 2, \ldots,$$

is a martingale with respect to $\{(\tau_n, \xi_n)\}$. We need only check [cf. Example (b) of Section 1] that

$$E\left[T_n - \frac{1}{aS_{n-1}} \,\middle|\, \tau_0, \ldots, \tau_{n-1}, \xi_0, \ldots, \xi_{n-1} \right] = 0. \tag{6.3}$$

The memoryless character of the exponential distribution and the definitions involved imply that T_n is the minimum of S_{n-1} independent lifetimes, which are all exponentially distributed with parameter a, and thus T_n is itself exponentially distributed with parameter aS_{n-1} (cf. Elementary Problem 1, Chapter 4). Equation (6.3) immediately ensues from these considerations.

The next calculation also exploits the martingale character of the sum and the exponential distribution underlying each summand. We get $Y_0 = 0$, and then

$$\begin{aligned} E[Y_n^2] &= \sum_{k=1}^{n} E[(Y_k - Y_{k-1})^2] \qquad \text{(see Problem 3)} \\ &= \sum_{k=1}^{n} E\left[\left(T_k - \frac{1}{aS_{k-1}} \right)^2 \right] \\ &= \sum_{k=1}^{n} E\left\{ E\left[\left(T_k - \frac{1}{aS_{k-1}} \right)^2 \middle| S_{k-1} \right] \right\} \end{aligned}$$

$$= \sum_{k=1}^{n} E\left[\left(\frac{1}{aS_{k-1}}\right)^2\right]$$

$$\leq \frac{1}{a^2}\left\{1 + \sum_{k=1}^{\infty} \frac{1}{k^2} E\left[\left(\frac{k}{S_k}\right)^2\right]\right\}.$$

In a moment we will show that $E[(k/S_k)^2]$ is uniformly bounded, say by C, and then

$$E[Y_n^2] \leq a^{-2}\left[1 + C\sum_{k=1}^{\infty} k^{-2}\right] < \infty, \qquad n = 1, 2, \ldots,$$

obtains, thereby verifying the conditions of the martingale mean square convergence theorem.

To obtain the bound C, let $V_k = \tilde{\xi}_1 + \tilde{\xi}_2 + \cdots + \tilde{\xi}_k$, where

$$\tilde{\xi}_i = \begin{cases} 0, & \text{if } \xi_i = 0, \\ 1, & \text{if } \xi_i \geq 1. \end{cases}$$

Then V_k has a binomial distribution with parameters k and $p = \Pr\{\xi_i > 0\} > 0$, for which we know

$$E[s^{V_k}] = [1 - p + ps]^k, \qquad 0 \leq s \leq 1,$$

and $1 + V_k \leq S_k$, whence

$$E\left[\frac{k}{S_k}\right] \leq kE\left[\frac{1}{1 + V_k}\right]$$

$$= k\int_0^1 E[s^{V_k}]\, ds$$

$$= k\int_0^1 [1 - p + ps]^k\, ds$$

$$= [k/(k+1)][1 - (1-p)^{k+1}]/p$$

$$\leq p^{-1} < \infty.$$

Note also that for $0 < c \leq 1$,

$$E\left[\frac{k}{c + \xi_1 + \cdots + \xi_k}\right] \leq \frac{1}{c} E\left[\frac{k}{1 + \xi_1 + \cdots + \xi_k}\right]$$

$$\leq (cp)^{-1} < \infty.$$

Next, using the elementary inequality $(a+b)^2 \geq 4ab$, together with the independence of $\xi_1, \xi_2, \ldots$, we obtain

$$E\left[\left(\frac{2k}{S_{2k}}\right)^2\right] = E\left[\left\{\frac{2k}{(\frac{1}{2}+\xi_1+\cdots+\xi_k)+(\frac{1}{2}+\xi_{k+1}+\cdots+\xi_{2k})}\right\}^2\right]$$
$$\leq E\left[\frac{k}{\frac{1}{2}+\xi_1+\cdots+\xi_k}\cdot\frac{k}{\frac{1}{2}+\xi_{k+1}+\cdots+\xi_{2k}}\right]$$
$$\leq \left\{E\left[\frac{k}{\frac{1}{2}+\xi_1+\cdots+\xi_k}\right]\right\}^2 \leq \left(\frac{2}{p}\right)^2.$$

This bounds $E[(n/S_n)^2]$ when n is even. When n is odd,

$$E\left[\left(\frac{n+1}{S_{n+1}}\right)^2\right] \leq \left(\frac{n+1}{n}\right)^2 E\left[\left(\frac{n}{S_n}\right)^2\right]$$
$$\leq 4\left(\frac{2}{p}\right)^2.$$

Thus $E[(k/S_k)^2] \leq C$ for all k if we take $C = 16/p^2$.

We apply the martingale mean square convergence theorem to conclude $Y_\infty = \lim_{n\to\infty} Y_n$ exists, with probability one. Since $\sum_{i=1}^{n} T_i = \tau_n$,

$$\tau_n - a^{-1}\sum_{i=1}^{n} S_{i-1}^{-1} \to Y_\infty, \qquad \text{as} \quad n\to\infty. \tag{6.4}$$

Further consequences can be drawn after we analyze the behavior of the series $\sum_{i=1}^{n} 1/(S_{i-1})$. Let $\mu = E[\xi_i] > 0$. From the strong law of large numbers we know that

$$\frac{1}{n}S_n \to \mu, \qquad \text{as} \quad n\to\infty,$$

with probability one. We write

$$\frac{1}{\log n}\sum_{i=1}^{n}\frac{1}{S_{i-1}} = \frac{1}{\log n}\sum_{i=1}^{N}\frac{1}{S_{i-1}} + \frac{1}{\log n}\sum_{i=N}^{n-1}\frac{1}{i}\frac{i}{S_i}.$$

For any $\varepsilon > 0$ we choose N so large that $|i/S_i - \mu^{-1}| \leq \varepsilon$ for all $i \geq N$. On the other hand, for any fixed N, the first term on the right goes to zero as $n\to\infty$. These estimates lead to the result

$$\limsup_{n\to\infty}\frac{1}{\log n}\left|\sum_{i=1}^{n}\frac{1}{S_{i-1}} - \sum_{i=N}^{n}\frac{1}{\mu i}\right| \leq \varepsilon,$$

and since ε is arbitrary, we infer

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=1}^{n} \frac{1}{S_{i-1}} = \frac{1}{\mu}, \tag{6.5}$$

with probability one. Combining the limits (6.4) and (6.5) we find that

$$\tau_n - \frac{\log n}{a\mu} \to Y_\infty, \qquad \text{as} \quad n \to \infty.$$

Again, the limit is in the probability one sense. In contrast, it appears remarkable that

$$\tau_{2n} - \tau_n \cong (a\mu)^{-1} \log 2$$

is asymptotically constant.

(f) *Doob's Process.* We will show that Doob's process [Example (k) of Section 1] is uniformly integrable and thus satisfies the full conditions of the basic martingale convergence theorem. Let Z, Y_0, Y_1, ... be joint random variables with $E[|Z|] < \infty$. We have shown (see p. 246) that

$$X_n = E[Z|Y_0, \ldots, Y_n], \qquad n = 0, 1, \ldots,$$

determines a martingale satisfying

$$E[|X_n|] \le E[|Z|], \qquad \text{for all} \quad n. \tag{6.6}$$

The maximal inequality for martingales yields that

$$\Pr\left\{\max_{0\le k\le n} |X_k| > \lambda\right\} \le \lambda^{-1} E[|X_n|] \le \lambda^{-1} E[|Z|],$$

and thus for $U = \sup_{k\ge 0} |X_k|$

$$\Pr\{U > \lambda\} \le \lambda^{-1} E[|Z|].$$

What is important is that

$$\lim_{\lambda\to\infty} \Pr\{U > \lambda\} = 0, \tag{6.7}$$

establishing that U is a finite-valued random variable. Now as $N \to \infty$,

$$\begin{aligned} E[|Z|] &\ge E[|Z| I\{0 \le U < N\}] \qquad \text{(concerning the notation } I\{\ \} \text{ see p. 279)} \\ &= \sum_{k=0}^{N-1} E[|Z| | k \le U < k+1] \Pr\{k \le U < k+1\} \\ &\to \sum_{k=0}^{\infty} E[|Z| | k \le U < k+1] \Pr\{k \le U < k+1\} \\ &= E[|Z|], \end{aligned}$$

implying that

$$\lim_{N\to\infty} E[|Z|I\{U \geq N\}] = 0. \tag{6.8}$$

We next verify the uniform integrability requirement of Remark 5.3. Consider

$$\begin{aligned} |E[X_n I\{|X_n| > c\}]| &\leq E[(I\{|X_n| > c\})(E[|Z||Y_0, \ldots, Y_n])] \\ &\leq E[|Z|I\{|X_n| > c\}] \\ &\leq E[|Z|I\{U > c\}] \to 0, \qquad \text{as} \quad c \to \infty, \end{aligned}$$

by (6.8). Thus, $\{X_n\}$ is uniformly integrable, and we may conclude that

$$\lim_{n\to\infty} X_n = X_\infty,$$

for some random variable X_∞. Moreover, the relations

$$E[X_\infty] = E[X_0]$$

and

$$E[|X_n - X_\infty|] \to 0, \qquad \text{as} \quad n \to \infty,$$

also prevail. It is correct that

$$X_\infty = E[Z|Y_0, Y_1, \ldots], \tag{6.9}$$

although the exact meaning of the conditional expectation in (6.9) is beyond our present scope. (See Section 7.)

For an application relevant to mathematical analysis, suppose W is a uniformly distributed random variable on [0, 1), and define Y_n by

$$Y_n = k2^{-n}, \qquad \text{for} \quad k2^{-n} \leq W < (k+1)2^{-n}$$

As noted in Example (1) of Section 1, $Y_0, \ldots, Y_n$ determine the first n bits in the terminating binary expansion of W.

Let f be an arbitrary function defined on [0, 1], for which

$$\int_0^1 |f(w)|\, dw < \infty.$$

Set $Z = f(W)$ and observe that

$$\begin{aligned} X_n &= E[Z|Y_0, \ldots, Y_n] \\ &= 2^n \int_{k/2^n}^{(k+1)/2^n} f(w)\, dw, \qquad \text{if} \quad k/2^n \leq W < (k+1)/2^n. \end{aligned}$$

From what we just proved about Doob's process, $\lim_{n\to\infty} X_n = X_\infty$ exists, and, since Y_0, Y_1, Y_2, ... give the full binary expansion of W, it is natural to suppose

$$\begin{aligned} X_\infty &= E[Z|\,Y_0,\ Y_1,\ \ldots] \\ &= E[f(W)|W] \\ &= f(W). \end{aligned}$$

While this is well beyond our scope, it is indeed valid, provided one adds the qualification "with probability one." We have then shown, with probability one,

$$f(W) = \lim_{n\to\infty} f_n(W),$$

where

$$f_n(w) = 2^n \int_{k(w)/2^n}^{[k(w)+1]/2^n} f(z)\,dz,$$

in which $k(w)$ is determined uniquely by the inequalities

$$k(w)/2^n \le w < [k(w)+1]/2^n.$$

Each approximating function f_n is a step function, constant on each interval $[k/2^n, (k+1)/2^n)$. Thus we have shown that an arbitrary integrable function f can be approximated by a sequence of step functions f_n in the sense that $f(w) = \lim_{n\to\infty} f_n(w)$ for "almost every" w, i.e., for every w in a set having probability one.

Finally the convergence $E[|f(W) - f_n(W)|] \to 0$ gives

$$\lim_{n\to\infty} \int_0^1 |f(z) - f_n(z)|\,dz = 0.$$

7: *Martingales with Respect to σ-Fields*

Until now we have always considered conditional expectations to be expectations computed under conditional distributions. This is mostly satisfactory for expressions of the form $E[X|\,Y_0, \ldots, Y_n]$, where X, Y_0, ..., Y_n possess a joint continuous density or are jointly discrete random variables. However, the analysis extended to the more complex expressions like $E[X|\,Y_0, Y_1, \ldots]$ or $E[X|\,Y(u), 0 \le u \le t]$ becomes more delicate.

The alternative and more modern approach is to define and evaluate conditional expectation, not with respect to a finite family of random variables, as we have done so far, but with respect to certain collections,

called σ-fields, of events. This suggests in a natural way a definition of a martingale with respect to a sequence of σ-fields. We will now sketch this formulation, so pervasive in contemporary writing.

REVIEW OF AXIOMATIC PROBABILITY THEORY

For the most part, this book has studied random variables only through their distributions. For example, at the very outset we considered a stochastic process $\{X(t); t \in T\}$ to be defined once all the finite-dimensional distributions

$$F(x_1, \ldots, x_n, t_1, \ldots, t_n) = \Pr\{X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n\}$$

were specified. A little more precision and structure is now needed.

Recall that the basic elements of probability theory are:

(1) The *sample space*, a set Ω whose elements ω correspond to the possible outcomes of an experiment;

(2) The *family of events*, a collection $\mathcal{F}$ of subsets A of Ω. We say that the event A *occurs* if the outcome ω of the experiment is an element of A; and

(3) The *probability measure*, a function P defined on $\mathcal{F}$ and satisfying

$$\text{(a)} \quad 0 = P[\varnothing] \leq P[A] \leq P[\Omega] = 1, \qquad \text{for} \quad A \in \mathcal{F} \quad (\varnothing = \text{the empty set}),$$

$$\text{(b)} \quad P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2], \qquad \text{for} \quad A_i \in \mathcal{F}, \quad i = 1, 2, \tag{7.1}$$

and

$$\text{(c)} \quad P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} P[A_n],$$

if $A_n \in \mathcal{F}$, are mutually disjoint ($A_i \cap A_j = \varnothing$, $i \neq j$).

The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Example. When there are only a denumerable number of possible outcomes, say $\Omega = \{\omega_1, \omega_2, \ldots\}$, we may take $\mathcal{F}$ to be the collection of all subsets of Ω. If $p_1, p_2, \ldots$ are nonnegative numbers with $\sum_n p_n = 1$, the assignment

$$P[A] = \sum_{\omega_i \in A} p_i$$

determines a probability measure defined on $\mathcal{F}$.

It is not always desirable, consistent, or feasible to take the family of events as the collection of *all* subsets of Ω. Indeed, when Ω is non-denumerably infinite, it may not be possible to define a probability measure on the collection of all subsets maintaining the properties of (7.1). In whatever way we prescribe $\mathscr{F}$ such that (7.1a)–(7.1c) hold, the family of events $\mathscr{F}$ should satisfy

$$\begin{aligned}
&\text{(a)} \quad \varnothing \in \mathscr{F},\ \Omega \in \mathscr{F};\\
&\text{(b)} \quad A^c \in \mathscr{F}, \text{ whenever } A \in \mathscr{F}, \text{ where } A^c = \{\omega \in \Omega;\ \omega \notin A\} \text{ is the}\\
&\qquad \text{complement of } A; \text{ and}\\
&\text{(c)} \quad \bigcup_{n=1}^{\infty} A_n \in \mathscr{F}, \text{ whenever } A_n \in \mathscr{F},\ n = 1, 2, \ldots.
\end{aligned} \tag{7.2}$$

A collection $\mathscr{F}$ of subsets of a set Ω satisfying (7.2a)–(7.2c) is called a *σ-field*. If $\mathscr{F}$ is a *σ-field*, then

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c\right)^c \in \mathscr{F},$$

whenever $A_n \in \mathscr{F}$, $n = 1, 2, \ldots$. Manifestly, as a consequence we find that finite unions and finite intersections of members of $\mathscr{F}$ are maintained in $\mathscr{F}$.

In this framework, a real random variable X is a real-valued function defined on Ω fulfilling certain "measurability" conditions given below. The distribution function of the random variable X is formally given by

$$\Pr\{a < X \leq b\} = P[\{\omega : a < X(\omega) \leq b\}]. \tag{7.3}$$

In words, the probability that the random variable X takes a value in $(a, b]$ is calculated as the probability of the set of outcomes ω for which $a < X(\omega) \leq b$. If relation (7.3) is to have meaning, X cannot be an arbitrary function on Ω, but must satisfy the condition that

$$\{\omega : a < X(\omega) \leq b\} \in \mathscr{F}, \qquad \text{for all real} \quad a < b,$$

since $\mathscr{F}$ embodies the only sets A for which $P[A]$ is defined. In fact, by exploiting the properties (7.2a)–(7.2c) of the σ-field $\mathscr{F}$, it is enough to require

$$\{\omega : X(\omega) \leq x\} \in \mathscr{F}, \qquad \text{for all} \quad x.$$

Let $\mathscr{A}$ be any σ-field of subsets of Ω. We say that X is *measurable with respect to $\mathscr{A}$*, or more briefly $\mathscr{A}$-measurable, if

$$\{\omega : X(\omega) \leq x\} \in \mathscr{A} \qquad \text{for all real} \quad x.$$

Thus, every real random variable is by definition $\mathscr{F}$-measurable. There may, in general, be smaller σ-fields with respect to which X is also measurable.

The σ-field *generated* by a random variable X is defined to be the smallest σ-field with respect to which X is measurable. It is denoted by $\mathcal{F}(X)$ and consists exactly of those sets A that are in every σ-field $\mathcal{A}$ for which X is $\mathcal{A}$-measurable. For example, if X has only denumerably many possible values $x_1, x_2, \ldots$ the sets

$$A_i = \{\omega : X(\omega) = x_i\} \qquad i = 1, 2, \ldots$$

form a countable *partition* of Ω, i.e.,

$$\Omega = \bigcup_{i=1}^{\infty} A_i,$$

and

$$A_i \cap A_j = \varnothing \qquad \text{if} \quad i \neq j,$$

and then $\mathcal{F}(X)$ includes precisely $\varnothing$, Ω, and every set that is the union of some of the A_i's.

Example. For the reader completely unfamiliar with this framework, the following simple example will help set the concepts. The experiment consists in tossing a nickel and a dime and observing "heads" or "tails." We take Ω to be

$$\Omega = \{(\mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T})\},$$

where for example, (H, T) stand for the outcome "nickel = heads, and dime = tails." We will take the collection of all subsets of Ω as the family of events. Assuming each outcome in Ω to be equally likely, we arrive at the probability measure:

$A \in \mathcal{F}$	$P[A]$	$A \in \mathcal{F}$	$P[A]$
ϕ	0	Ω	1
{(H, H)}	$\frac{1}{4}$	{(H, T), (T, H), (T, T)}	$\frac{3}{4}$
{(H, T)}	$\frac{1}{4}$	{(H, H), (T, H), (T, T)}	$\frac{3}{4}$
{(T, H)}	$\frac{1}{4}$	{(H, H), (H, T), (T, T)}	$\frac{3}{4}$
{(T, T)}	$\frac{1}{4}$	{(H, H), (H, T), (T, H)}	$\frac{3}{4}$
{(H, H), (H, T)}	$\frac{1}{2}$	{(T, H), (T, T)}	$\frac{1}{2}$
{(H, H), (T, H)}	$\frac{1}{2}$	{(H, T), (T, T)}	$\frac{1}{2}$
{(H, H), (T, T)}	$\frac{1}{2}$	{(H, T), (T, H)}	$\frac{1}{2}$

The event "the nickel is heads" is {(H, H), (H, T)} and has, according to the table, probability $\frac{1}{2}$, as it should.

Let X_n be 1 if the nickel is heads, and 0 otherwise, let X_d be the corresponding random variable for the dime, and let $Z = X_n + X_d$ be the total number of heads. As functions on Ω, we have

$\omega \in \Omega$	$X_n(\omega)$	$X_d(\omega)$	$Z(\omega)$
(H, H)	1	1	2
(H, T)	1	0	1
(T, H)	0	1	1
(T, T)	0	0	0

Finally, the σ-fields generated by X_n and Z are:

$$\mathscr{F}(X_n) = \phi, \Omega, \{(H, H), (H, T)\}, \{(T, H), (T, T)\},$$

and

$$\begin{aligned}\mathscr{F}(Z) = \phi, \Omega, &\{(H, H)\}, \{(H, T), (T, H)\}, \{(T, T)\},\\ &\{(H, T), (T, H), (T, T)\}, \{(H, H), (T, T)\},\\ &\{(H, H), (H, T), (T, H)\}.\end{aligned}$$

$\mathscr{F}(X_n)$ contains 4 sets and $\mathscr{F}(Z)$ contains 8. Is X_n measurable with respect to $\mathscr{F}(Z)$, or vice versa?

Every pair X, Y of random variables determines a σ-field called the σ-field generated by X, Y. It is the smallest σ-field with respect to which both X and Y are measurable. This field comprises exactly those sets A that are in every σ-field $\mathscr{A}$ for which X and Y are both $\mathscr{A}$-measurable. If both X and Y assume only denumerably many possible values, say $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$, respectively, then the sets

$$A_{ij} = \{\omega : X(\omega) = x_i, Y(\omega) = y_j\}, \qquad i, j = 1, 2, \ldots,$$

present a countable partition of Ω and $\mathscr{F}(X, Y)$ consists precisely of ϕ, Ω, and every set that is the union of some of the A_{ij}'s. Observe that X is measurable with respect to $\mathscr{F}(X, Y)$, and thus $\mathscr{F}(X) \subset \mathscr{F}(X, Y)$.

More generally, let $\{X(t); t \in T\}$ be any family of random variables. Then the σ-field generated by $\{X(t); t \in T\}$ is the smallest σ-field with respect to which every random variable $X(t)$, $t \in T$, is measurable. It is denoted by $\mathscr{F}\{X(t); t \in T\}$.

A special role is played by a distinguished σ-field of sets of real numbers. The σ-field of *Borel sets* is the σ-field generated by the identity function $f(x) = x$, for $x \in (-\infty, \infty)$. Alternatively, the σ-field of Borel sets is the smallest σ-field containing every interval of the form $(a, b]$, $-\infty \leq a \leq b < +\infty$. A real-valued function of a real variable is said to be *Borel measurable* if it is measurable with respect to the σ-field of Borel sets.

In n-dimensional Euclidean space, the σ-field of Borel sets is the σ-field generated by the set of functions

$$\{f_i(x_1, \ldots, x_n) = x_i;\ i = 1, \ldots, n\}.$$

CONDITIONAL EXPECTATION WITH RESPECT TO A σ-FIELD

While every random variable Y generates a σ-field $\mathcal{F}(Y)$, it is not true that every σ-field arises in this manner. Thus, the concept of conditional expectation with respect to a σ-field, our next topic, is a strict extension of the concept of conditional expectation with respect to random variables.

We begin with the formal definition.

Definition 7.1. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ for which $E[|X|] < \infty$. Let $\mathcal{B}$ be a σ-field contained in $\mathcal{F}$, i.e., every set $B \in \mathcal{B}$ is also a member of $\mathcal{F}$. The conditional expectation of X with respect to $\mathcal{B}$ is defined to be any random variable $E[X|\mathcal{B}]$ having the properties:*

(i) *$E[X|\mathcal{B}]$ is a measurable function with respect to $\mathcal{B}$; and*

(ii) *$E[XI_B] = E\{E[X|\mathcal{B}]I_B\}$, for all $B \in \mathcal{B}$, where I_B is the indicator function for the event B. Alternatively,* (ii) *may be replaced by the equivalent*

(ii′) *$E[XZ] = E\{E[X|\mathcal{B}]Z\}$ for every bounded random variable Z that is $\mathcal{B}$-measurable.*

Several remarks are in order. First, it can be shown that a random variable $E[X|\mathcal{B}]$ satisfying (i) and (ii) exists whenever $E[|X|] < \infty$, so that the definition is nonvoid. In fact, a meaningful definition results as long as not both $E[X^+]$ and $E[X^-]$ are infinite. On the other hand, the definition is ambiguous in that there may be more than one conditional expectation $E[X|\mathcal{B}]$ satisfying (i) and (ii). Thus we speak of different "versions" of the conditional expectation. Fortunately, any two versions are equal, with probability one. That is, if $E^{(1)}[X|\mathcal{B}]$ and $E^{(2)}[X|\mathcal{B}]$ satisfy (i) and (ii), then

$$P\{E^{(1)}[X|\mathcal{B}] = E^{(2)}[X|\mathcal{B}]\} = 1.$$

Thus, in probability terms, the ambiguity causes no difficulty. Finally, let us mention the equivalence of (ii) and (ii′). Clearly (ii′) implies (ii), since we may always take Z to be the bounded $\mathcal{B}$-measurable function

$$Z(\omega) = I_B(\omega) = \begin{cases} 1, & \text{if } \omega \in B, \\ 0, & \text{if } \omega \notin B. \end{cases}$$

The converse implication is validated by suitably approximating an arbitrary Z by step functions of the form

$$Z_n(\omega) = \sum_{i=1}^{n} \alpha_{in} 1_{B_i}(\omega),$$

where $\{B_1, \ldots, B_n\}$ is a finite partition of Ω with $B_i \in \mathscr{B}$. Then (ii′) holds for such Z_n, and by passing to the limit, we can infer (ii′) for arbitrary bounded $\mathscr{B}$-measurable random variables.

In Chapter 1 we expressed the conditional expectation of X given a random variable $Y = y$ as

(I) Any function $E[X|Y=y]$ of y which satisfies

(II) $E[Xg(Y)] = \int E[X|Y=y]g(y)\,dF_Y(y)$, for every bounded function g.

One would hope that our current definition extends the earlier one, that is, coincides with it when $\mathscr{B} = \mathscr{F}(Y)$ is the σ-field generated by the random variable Y. This is indeed the case. One can show that a random variable Z is measurable with respect to $\mathscr{B} = \mathscr{F}(Y)$ if and only if one can write $Z = f(Y)$ for some Borel measurable function f. (Problem 13 asks for a proof of this statement when $\mathscr{B}$ is the σ-field of unions of some countable partition $\mathscr{B}_0$.) Then (I) states that $E[X|Y]$ is measurable with respect to $\mathscr{B} = \mathscr{F}(Y)$ and (II) yields (ii′), $E[XZ] = E\{E[X|Y]Z\}$ for all bounded $Z = f(Y)$. Thereby, Definition 7.1 extends the concept of conditional expectation with respect to a random variable.

Last, let us work out what Definition 7.1 means in the case that $\Omega = \{\omega_1, \omega_2, \ldots\}$ is denumerable, $\mathscr{F}$ consists of all subsets of Ω, and P is evaluated by the formula

$$P[A] = \sum_{\omega_i \in A} p_i, \qquad A \in \mathscr{F}, \tag{7.4}$$

where $p_1, p_2, \ldots$ are nonnegative numbers summing to one. The expectation of a random variable X is defined by

$$E[X] = \sum_{j=1}^{\infty} X(\omega_j)p_j,$$

provided the sum converges absolutely. Let $\mathscr{B}_0 = \{B_1, B_2, \ldots\}$ be a denumerable partition of Ω, and let $\mathscr{B}$ be the σ-field consisting of ϕ, Ω and all possible unions of sets in $\mathscr{B}_0$. For each B_j in $\mathscr{B}_0$, define the elementary conditional expectation

$$E[X|B_j] = \sum_{\omega_k \in B_j} X(\omega_k)P[\{\omega_k\}|B_j],$$

where

$$P[A|B_j] = P[A \cap B_j]/P[B_j], \qquad A \in \mathscr{F}. \tag{7.5}$$

The definition breaks down when $P[B_j] = 0$, since then the right-hand side of (7.5) is 0/0. Arbitrarily set $E[X|B_j] = 17$ whenever $P[B_j] = 0$, which completes the definition and, incidentally, indicates where the lack of uniqueness in Definition 7.1 arises. The next step is to make the collection of numbers $E[X|B_j]_{j=1}^{\infty}$ into a random variable $E[X|\mathscr{B}]$ by the formula

$$\begin{aligned} E[X|\mathscr{B}](\omega) &= E[X|B_j], \qquad \text{if} \quad \omega \in B_j, \\ &= \sum_{j=1}^{\infty} E[X|B_j] I_{B_j}(\omega), \end{aligned}$$

where

$$I_{B_j}(\omega) = \begin{cases} 1, & \text{if} \quad \omega \in B_j, \\ 0, & \text{if} \quad \omega \notin B_j. \end{cases}$$

Then, the random variable $E[X|\mathscr{B}]$, being constant on each of the sets B_j, is $\mathscr{B}$-measurable. We check (ii) of Definition 7.1. Let $B \in \mathscr{B}$ be prescribed, say $B = \bigcup_{n=1}^{\infty} B_n'$, where $B_n' \in \mathscr{B}_0$, $n = 1, 2 \ldots$. On the one hand,

$$E[XI_B] = \sum_{n=1}^{\infty} \sum_{\omega_k \in B_n'} X(\omega_k) p_k,$$

while at the same time,

$$\begin{aligned} E\{E[X|\mathscr{B}] I_B\} &= \sum_{n=1}^{\infty} E[X|B_n'] P[B_n'] \\ &= \sum_{n=1}^{\infty} \sum_{\omega_k \in B_n'} X(\omega_k) p_k. \end{aligned}$$

The equality of the two right-hand sides verifies (ii).

Let us show that the elementary definition for $E[X|B_j]$ can be recovered from Definition 7.1 by taking $B = B_j$, provided $P[B_j] > 0$. Then (ii) becomes

$$\sum_{\omega_k \in B_j} X(\omega_k) p_k = \sum_{\omega_k \in B_j} E[X|\mathscr{B}](\omega_k) p_k. \tag{7.6}$$

Now $E[X|\mathscr{B}]$ is $\mathscr{B}$-measurable if and only if it has a constant value, say a_j, on each of the sets B_j. Then (7.6) becomes

$$\sum_{\omega_k \in B_j} X(\omega_k) p_k = a_j P[B_j],$$

and the constant value a_j must be

$$\begin{aligned} a_j &= \sum_{\omega_k \in B_j} X(\omega_k) p_k / P[B_j] \\ &= E[X|B_j], \end{aligned}$$

whenever $P[B_j] > 0$.

Thus, one can recover our intuitive concept of conditional expectation from Definition 7.1 whenever $\mathscr{B}$ is the σ-field of unions of a countable partition $\mathscr{B}_0$.

The definition of conditional expectation given a random variable was expectation computed under a conditional distribution. Conditional expectations with respect to σ-fields inherit most of the familiar properties, provided we interpret "equal" as meaning "equal with probability one." For reference, we list some of these properties, corresponding to (1.6), (1.7), and (1.10)–(1.14) of Chapter 1. There is no satisfactory analog of (1.8). We suppose X, X_1, and X_2 are random variables having finite expectations, a_1, a_2 real numbers, and $\mathscr{A}$ and $\mathscr{B}$ are sub-σ-fields of $\mathscr{F}$. Then

$$E[a_1X_1 + a_2X_2|\mathscr{B}] = a_1E[X_1|\mathscr{B}] + a_2E[X_2|\mathscr{B}]; \tag{7.7}$$

$$X \geq 0 \quad \text{implies} \quad E[X|\mathscr{B}] \geq 0; \tag{7.8}$$

$$E[XZ|\mathscr{B}] = ZE[X|\mathscr{B}], \quad \text{for every bounded } \mathscr{B}\text{-measurable } Z; \tag{7.9}$$

$$E[XZ] = E\{Z \cdot E[X|\mathscr{B}]\}, \quad \text{for every bounded } \mathscr{B}\text{-measurable } Z; \tag{7.10}$$

$$E[Z|\mathscr{B}] = Z, \quad \text{for every } \mathscr{B}\text{-measurable } Z \text{ satisfying } E[|Z|] < \infty; \tag{7.11}$$

$$E[X|\mathscr{A}] = E\{E[X|\mathscr{B}]|\mathscr{A}\}, \quad \text{if } \mathscr{A} \subset \mathscr{B}; \tag{7.12}$$

$$E[X] = E\{E[X|\mathscr{B}]\} \quad \text{(the law of total probability)}. \tag{7.13}$$

The style of proof used in validating these properties is typified by examining (7.7). We show that $a_1E[X_1|\mathscr{B}] + a_2E[X_2|\mathscr{B}]$ satisfies the defining conditions for $E[a_1X_1 + a_2X_2|\mathscr{B}]$. First, it can be checked that a linear combination of two $\mathscr{B}$-measurable random variables is also $\mathscr{B}$-measurable. (Problem 14 requests a proof of this statement, where $\mathscr{B}$ is the σ-fields of unions of a countable partition $\mathscr{B}_0$.) Thus, $a_1E[X_1|\mathscr{B}] + a_2E[X_2|\mathscr{B}]$ satisfies stipulation (i) of Definition 7.1. Substitute $a_1E[X_1|\mathscr{B}] + a_2E[X_2|\mathscr{B}]$ into what is required, for $E[a_1X_1 + a_2X_2|\mathscr{B}]$ in condition (ii′) and see if it is satisfied. Consider any bounded $\mathscr{B}$-measurable Z. The justification of the succeeding steps is routine:

$$\begin{aligned} E[\{a_1X_1 + a_2X_2\}Z] &= a_1E[X_1Z] + a_2E[X_2Z] \\ &= a_1E\{E[X_1|\mathscr{B}]Z\} + a_2E\{E[X_2|\mathscr{B}]Z\} \\ &= E\{(a_1E[X_1|\mathscr{B}] + a_2E[X_2|\mathscr{B}])Z\}. \end{aligned}$$

This completes the validation of (7.7).

MARTINGALES WITH RESPECT TO AN INCREASING FAMILY OF σ-FIELDS

Let $\{X_n\}_0^\infty$ be a sequence of real random variables on a probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_n\}$ be a sequence of sub-σ-fields of $\mathcal{F}$ with

$$\mathcal{F}_0 \subset \mathcal{F}_1 \cdots \subset \mathcal{F}_n \subset \cdots \subset \mathcal{F}.$$

We say that $\{X_n\}$ is adapted to $\{\mathcal{F}_n\}$ if, for every n, X_n *is $\mathcal{F}_n$-measurable.* For example, suppose $Y_0, Y_1, \ldots$ are also defined on $(\Omega, \mathcal{F}, P)$ and $\mathcal{F}_n$ is the σ-field generated by $\{Y_0, Y_1, \ldots, Y_n\}$. Then

$$\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}, \qquad \text{for all} \quad n,$$

and if $X_n = g_n(Y_0, \ldots, Y_n)$ for a sequence of Borel measurable functions $g_n(\cdot, \ldots, \cdot)$, then $\{X_n\}$ is adapted to $\{\mathcal{F}_n\}$.

We again think of $\mathcal{F}_n$ as containing the information available at stage n, just as we did earlier with $(Y_0, \ldots, Y_n)$. Then X_n is measurable with respect to $\mathcal{F}_n = \mathcal{F}(Y_0, \ldots, Y_n)$ if and only if, in our earlier usage, X_n is *determined* by $(Y_0, \ldots, Y_n)$.

The relation $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ expresses the increase in information as n progresses.

Definition 7.2. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_n\}$ be a sequence of sub-σ-fields of $\mathcal{F}$ with*

$$\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}, \qquad \textit{for all} \quad n.$$

Then $\{X_n\}$ is called a submartingale with respect to $\{\mathcal{F}_n\}$ if:

(i) *$\{X_n\}$ is adapted to $\{\mathcal{F}_n\}$ (that is each X_n is $\mathcal{F}_n$-measurable),*
(ii) *$E[X_n^+] < \infty$, for all n, and*
(iii) *$E[X_{n+1}|\mathcal{F}_n] \geq X_n$, for all n.*

If $\{-X_n\}$ is a submartingale, then $\{X_n\}$ is called a supermartingale. If both $\{X_n\}$ and $\{-X_n\}$ are submartingales, then $\{X_n\}$ is called a martingale with respect to $\{\mathcal{F}_n\}$.

A few remarks may be helpful. If $\{X_n\}$ is a martingale, then

$$X_n = E[X_{n+1}|\mathcal{F}_n]. \tag{7.14}$$

The right-hand side, and hence the left, is $\mathcal{F}_n$-measurable by the definition of conditional expectation. Hence the representation (7.14) implies (i) for a martingale. Requirement (iii) can be stated in an equivalent form by using the properties of conditional expectation. We get

(iii′) $E[X_{n+1}Z] \geq E[X_n Z]$, for all bounded $\mathcal{F}_n$-measurable $Z \geq 0$.

To infer (iii') from (iii), use the definition of conditional expectation with respect to σ-fields and properties (7.7) and (7.8) to obtain

$$E[X_{n+1}Z] = E\{E[X_{n+1}|\mathscr{F}_n]Z\} \geq E[X_n Z].$$

To deduce (iii) from (iii'), use

$$E\{E[X_{n+1}|\mathscr{F}_n]Z\} = E[X_{n+1}Z]$$

to obtain

$$E\{(E[X_{n+1}|\mathscr{F}_n] - X_n)Z\} \geq 0, \qquad \text{for all bounded } \mathscr{F}_n\text{-measurable } Z \geq 0.$$

Now $Y = E[X_{n+1}|\mathscr{F}_n] - X_n$ is $\mathscr{F}_n$-measurable, and if $E[YZ] \geq 0$ for all bounded $\mathscr{F}_n$-measurable $Z \geq 0$, then $Y \geq 0$. (Problem 15 asks for a verification of this statement, where $\mathscr{B}$ is the σ-field consisting of unions of sets in a denumerable partition $\mathscr{B}_0$.) Of course, the relation

$$Y = E[X_{n+1}|\mathscr{F}_n] - X_n \geq 0$$

is the desired (iii).

All of the results concerning martingales with respect to random variables that were developed earlier in this chapter carry over to martingales with respect to increasing σ-fields, with only technical modifications required. More explicitly, once we produce a definition for a Markov time with respect to a sequence of σ-fields, the optional sampling theorems and the martingale convergence theorems will persist. We will not repeat the entire development, but only the early part of it in the martingale case, and this mainly for the pedagogical practice it provides in manipulating conditional expectations with respect to σ-fields.

Proposition 7.1. *Let $\{X_n\}$ be a martingale with respect to $\{\mathscr{F}_n\}$. Then $E[X_n] = E[X_0]$ for all n.*

Proof. Applying the law of total probability (7.13) in the form $E[X] = E\{E[X|\mathscr{B}]\}$ to the martingale equality $X_n = E[X_{n+1}|\mathscr{F}_n]$, we obtain

$$\begin{aligned} E[X_n] &= E\{E[X_{n+1}|\mathscr{F}_n]\} \\ &= E[X_{n+1}]. \end{aligned}$$

An induction completes the proof. ■

Proposition 7.2. *Let $\{X_n\}$ be a martingale with respect to $\{\mathscr{F}_n\}$. If Z is a bounded $\mathscr{F}_n$-measurable random variable, then*

$$E[ZX_{n+k}|\mathscr{F}_n] = ZX_n, \qquad n = 0, 1, \ldots; \qquad k \geq 1$$

Proof. We appeal to property (7.9),

$$E[ZX_{n+k}|\mathscr{F}_n] = ZE[X_{n+k}|\mathscr{F}_n], \text{ for } Z \text{ bounded and } \mathscr{F}_n\text{-measurable.} \tag{7.15}$$

Now on the basis of (7.12), we obtain

$$\begin{aligned} E[X_{n+k}|\mathscr{F}_n] &= E\{E[X_{n+k}|\mathscr{F}_{n+k-1}]|\mathscr{F}_n\} \\ &= E[X_{n+k-1}|\mathscr{F}_n], \end{aligned}$$

which continues by induction until

$$E[X_{n+k}|\mathscr{F}_n] = E[X_{n+1}|\mathscr{F}_n] = X_n.$$

This together with (7.15) completes the proof. ■

Let $\{\mathscr{F}_n\}$ be a sequence of sub-σ-fields of $\mathscr{F}$ satisfying $\mathscr{F}_n \subset \mathscr{F}_{n+1} \subset \mathscr{F}$, for all n. *A random variable T taking values in* $\{0, 1, \ldots, \infty\}$ is called a *Markov time* with respect to $\{\mathscr{F}_n\}$, if for every $n = 0, 1, 2, \ldots$, the event $\{T = n\}$ is in $\mathscr{F}_n$. Recall that every random variable is a function defined on the sample space Ω. Thus we require

$$\{\omega : T(\omega) = n\} \in \mathscr{F}_n, \qquad \text{for all} \quad n. \tag{7.16}$$

Alternatively, using the facts that each $\mathscr{F}_n$ is a σ-field and $\mathscr{F}_n \subset \mathscr{F}_{n+1}$ for all n, requirement (7.16) may be replaced by either of the equivalent conditions

$$\{\omega : T(\omega) \leq n\} \in \mathscr{F}_n, \qquad \text{for all} \quad n, \tag{7.17}$$

or

$$\{\omega : T(\omega) > n\} \in \mathscr{F}_n, \qquad \text{for all} \quad n. \tag{7.18}$$

For example, to conclude (7.17) from (7.16), observe

$$\{\omega;\ T(\omega) \leq n\} = \bigcup_{k=0}^{n} \{\omega;\ T(\omega) = k\}.$$

Now each $\{\omega : T(\omega) = k\} \in \mathscr{F}_k \subset \mathscr{F}_n$, so the union belongs to $\mathscr{F}_n$, and (7.17) is satisfied.

Alternatively, we can require that every indicator random variable

$$I_{\{T=n\}} = \begin{cases} 1, & \text{if} \quad T = n, \\ 0 & \text{if} \quad T \neq n, \end{cases}$$

be measurable with respect to $\mathscr{F}_n$.

If $\mathscr{F}_n = \mathscr{F}(Y_0, Y_1, \ldots, Y_n)$ is the σ-field generated by the random variables $Y_0, \ldots, Y_n$, then $I_{\{T=n\}}$ is measurable with respect to $\mathscr{F}_n$ if and only if

$$I_{\{T=n\}} = g_n(Y_0, \ldots, Y_n),$$

for some appropriately measurable function $g_n(\cdot, \ldots, \cdot)$. Thus, the latest definition of Markov time is an extension of the earlier one.

Every constant $T \equiv n$ is a Markov time. If T and S are Markov times, so are $T + S$, $T \wedge S = \min\{T, S\}$, and $T \vee S = \max\{T, S\}$. A sample proof is

$$\{\omega : T \wedge S > n\} = \{\omega : T > n\} \cap \{\omega : S > n\} \in \mathscr{F}_n.$$

Lemma 7.1. *Let $\{X_n\}$ be a martingale and T a Markov time with respect to $\{\mathscr{F}_n\}$. Then for all $n \geq k$,*

$$E[X_n I_{\{T=k\}}] = E[X_k I_{\{T=k\}}].$$

($I_{\{\cdot\}}$ is the indicator function of the event described in $\{\cdot\}$.)

Proof. We use the fact that $I_{\{T=k\}}$ is $\mathscr{F}_k$-measurable. Then,

$$\begin{aligned} E[X_n I_{\{T=k\}}] &= E\{E[X_n I_{\{T=k\}} | \mathscr{F}_k]\} && \text{(by the law of total probability)} \\ &= E\{I_{\{T=k\}} E[X_n | \mathscr{F}_k]\} && \text{[by (7.10)]} \\ &= E[X_k I_{\{T=k\}}]. \quad \blacksquare \end{aligned}$$

From this point on, the study of martingales with respect to σ-fields paraphrases that carried out in Sections 1–6, for example, as in the following lemma.

Lemma 7.2. *Let $\{X_n\}$ be a martingale and T a Markov time with respect to $\{\mathscr{F}_n\}$. Then for all $n = 0, 1, \ldots,$*

$$E[X_0] = E[X_{T \wedge n}] = E[X_n].$$

Proof. As in Lemma 3.2, mutatis mutandis, relying on Lemma 7.1 rather than Lemma 3.1. $\blacksquare$

The crossings inequality, maximal inequality, optional sampling theorems, and martingale convergence theorems all follow from Lemma 7.2 *just as they did earlier from Lemma* 3.2. We proceed to an example.

Example. Let $\{Y_n\}$ be random variables on some probability space $(\Omega, \mathscr{F}, P)$, and for each n let $\mathscr{F}_n$ be the σ-field generated by $(Y_0, \ldots, Y_n)$. If Z satisfies $E[|Z|] < \infty$, we pointed out in Example (k) of Section 1 that

$$X_n = E[Z | \mathscr{F}_n], \qquad n = 0, 1, \ldots,$$

constitutes a martingale, and we established in Example (f) of Section 6 that this martingale was uniformly integrable. The martingale convergence theorem affirms that

$$\lim_{n\to\infty} X_n = X_\infty$$

exists, with probability one and, moreover,

$$\lim_{n\to\infty} E[|X_n - X_\infty|] = 0$$

holds. We mentioned that X_∞ could be represented as

$$X_\infty = E[Z|Y_0, Y_1, \ldots].$$

Now interpreting the right-hand side as an expectation of Z under a conditional distribution is rather delicate. However, if we prescribe $\mathcal{F}_\infty$ as the σ-field generated by $(Y_0, Y_1, \ldots)$ it is not hard to motivate the formula

$$X_\infty = E[Z|\mathcal{F}_\infty].$$

In accordance with Definition 7.1, we need to show first that X_∞ is $\mathcal{F}_\infty$-measurable and second that for every bounded $\mathcal{F}_\infty$-measurable random variable W, the equation

$$E[X_\infty W] = E[ZW] \tag{7.19}$$

obtains. Each X_n is $\mathcal{F}_n$-measurable, and hence $\mathcal{F}_\infty$-measurable, since $\mathcal{F}_n \subset \mathcal{F}_\infty$ for all n. It follows that $X_\infty = \lim_{n\to\infty} X_n$ is $\mathcal{F}_\infty$-measurable. Another way to view this is that each $X_n = E[Z|Y_0, \ldots, Y_n]$ is a function of $Y_0, \ldots, Y_n$, so that $X_\infty = \lim X_n$ is an appropriately measurable function of the entire sequence $Y_0, Y_1, \ldots$, and hence measurable with respect to $\mathcal{F}_\infty$.

To prove (7.19), it suffices to consider a bounded $\mathcal{F}_m$-measurable W_m for arbitary m. The general case follows by suitably approximating the $\mathcal{F}_\infty$-measurable W by random variables W_m with m increasing. But if W_m is $\mathcal{F}_m$-measurable,

$$\begin{aligned} E[X_n W_m] &= E\{E[Z|\mathcal{F}_n]W_m\} \\ &= E\{E[ZW_m|\mathcal{F}_n]\} \text{ if } n \geq m && \text{(since } W_m \text{ is } \mathcal{F}_m \subset \mathcal{F}_n \text{ measurable)} \\ &= E[ZW_m] && \text{(by the law of total probabilities).} \end{aligned}$$

Passing to the limit with n yields

$$E[X_\infty W_m] = \lim_{n\to\infty} E[X_n W_m] = E[ZW_m],$$

and (7.19) is proved.

We pause in the example to bring out an observation. Since $X_\infty = \lim_{n\to\infty} X_n$ is $\mathcal{F}_\infty$-measurable, we know from (7.11)

$$X_\infty = E[X_\infty|\mathcal{F}_\infty].$$

On the other hand, we have just validated the representation

$$X_\infty = E[Z|\mathcal{F}_\infty],$$

and by virtue of property (7.12) and the fact of $\mathcal{F}_n \subset \mathcal{F}_\infty$, we have

$$\begin{aligned} X_n &= E[Z|\mathcal{F}_n] \\ &= E\{E[Z|\mathcal{F}_\infty]|\mathcal{F}_n\} \\ &= E[X_\infty|\mathcal{F}_n]. \end{aligned}$$

That is, $X_n = E[X_\infty|\mathcal{F}_n]$, where $X_\infty = \lim_{n\to\infty} X_n$. That this is correct for every uniformly integrable martingale is worth highlighting as a lemma.

Lemma 7.3. *Let $\{X_n\}$ be a uniformly integrable martingale (see p. 258) with respect to $\{\mathcal{F}_n\}$. Then*

$$X_n = E[X_\infty|\mathcal{F}_n]$$

where

$$X_\infty = \lim_{n\to\infty} X_n.$$

Proof. The basic martingale convergence theorem guarantees the existence of $X_\infty = \lim_{n\to\infty} X_n$ and the fact of

$$\lim_{n\to\infty} E[|X_n - X_\infty|] = 0. \tag{7.20}$$

We now show that X_n possesses the properties required of $E[X_\infty|\mathcal{F}_n]$ in line with Definition 7.1. Note, first, since $\{X_n\}$ is a martingale that X_n is $\mathcal{F}_n$-measurable. Let W be a bounded $\mathcal{F}_n$-measurable random variable. Then

$$\begin{aligned} E[X_\infty W] &= E\left[\lim_{m\to\infty} X_m W\right] \\ &= \lim_{m\to\infty} E[X_m W] \qquad \text{(the justification of interchange of limit and expectation is given below)} \\ &= E[X_n W]. \end{aligned}$$

That is, X_n satisfies the requirements for $E[X_\infty|\mathcal{F}_n]$. The interchange of limits is legitimate in view of the inequalities

$$\begin{aligned}|E[X_\infty W] - E[X_n W]| &\leq E[|X_\infty W - X_n W|] \\ &\leq AE[|X_n - X_\infty|],\end{aligned}$$

where $A < \infty$ is such that $|W| \leq A$, and now appeal to (7.20). ∎

We have established the important result that *every* uniformly integrable martingale $\{X_n\}$ has the form of a Doob's process $X_n = E[Z|\mathcal{F}_n]$ for $Z = X_\infty = \lim_{n\to\infty} X_n$.

An Application to Mathematical Analysis. Let f be a real-valued function on $[0, 1]$ that is Lipschitz continuous, i.e., f satisfies

$$|f(x) - f(y)| \leq C|x - y|, \qquad \text{for all} \quad x, y \in [0, 1],$$

where $C < \infty$ is a constant. For $n = 1, 2, \ldots$, specify $\mathcal{P}_n$ as the partition of $[0, 1)$ given by

$$\mathcal{P}_n = \{[k/2^n, [k+1]/2^n); k = 0, \ldots, 2^n - 1\}.$$

Determine $\mathcal{F}_n$ as the σ-field consisting of ϕ, $\Omega = [0, 1)$ and unions of sets in $\mathcal{P}_n$.

Let Z have a uniform distribution on $[0, 1)$ and define the sequence

$$X_n = 2^n\{f(k/2^n) - f((k-1)/2^n)\}, \qquad \text{if} \quad (k-1)/2^n \leq Z < k/2^n.$$

Then $\{X_n\}$ is a martingale with respect to $\{\mathcal{F}_n\}$. In fact, $\mathcal{F}_n$ is the σ-field generated by $Y_0, \ldots, Y_n$, where

$$Y_n = k/2^n, \qquad \text{for } k \text{ satisfying} \quad k/2^n \leq Z < (k+1)/2^n,$$

and we verified the martingale property in Example (l) of Section 1.

Observe that X_n is approximately the derivative of f at the (randomly chosen) point Z. Of course, f may not be differentiable, but being Lipschitz continuous, $|X_n| \leq C$ for all n, hence $\{X_n\}$ is uniformly integrable, and consequently

$$X_\infty = \lim_{n\to\infty} X_n$$

exists for a set of $Z \in [0, 1)$ having probability one. By Lemma 7.3, we have

$$X_n = E[X_\infty|\mathcal{F}_n]. \tag{7.21}$$

We make explicit the fact that X_∞ is some function g of the random variable Z, by writing $X_\infty = g(Z)$.

Take $B = [0, k/2^n) \in \mathscr{F}_n$. Then from (7.21),

$$\begin{aligned} E[X_n I_B] &= f(k/2^n) - f(0) \\ &= E[X_\infty I_B] \\ &= \int_0^{k/2^n} g(x)\, dx. \end{aligned}$$

By passing to the limit in a sequence of binary rationals converging to an arbitrary $z \in [0, 1)$, it follows that

$$f(z) - f(0) = \int_0^z g(x)\, dx.$$

In this sense, g is a derivative of f, the so-called Radon–Nikodym derivative.

8: Other Martingales

The martingale concept requires only that the index set T of the process $\{X(t);\ t \in T\}$ have some notion of ordering. In particular, T may be any subset of the real line.

Definition 8.1. *Let T be a set in $(-\infty, +\infty)$, and let $\{X(t);\ t \in T\}$ be a stochastic process defined on a probability space $(\Omega, \mathscr{F}, P)$. For each $t \in T$, suppose $\mathscr{F}_t$ is a sub-σ-field of $\mathscr{F}$ and*

$$\mathscr{F}_t \subset \mathscr{F}_s, \qquad \text{if} \quad t < s, \quad t, s \in T.$$

Then $\{X(t)\}$ is called a submartingale with respect to $\{\mathscr{F}_t\}$ if for all $t \in T$,

(i) *$X(t)$ is $\mathscr{F}_t$-measurable,*
(ii) *$E[X(t)^+] < \infty$, and*
(iii) *$E[X(t+u)|\mathscr{F}_t] \geq X(t)$, $u > 0$, $t + u \in T$.*

To continue, $\{X(t)\}$ is called a supermartingale if $\{-X(t)\}$ is a submartingale, and a martingale if it is both a supermartingale and a submartingale.

A number of cases commonly arise:

$T = \{\ldots, -2, -1, 0\}$,	the set of negative integers,
$T = \{\ldots, -1, 0, 1, \ldots\}$,	the set of all integers,
$T = [0, \infty)$,	the positive real line,
$T = (-\infty, \infty)$,	the total real line,

and even

$T = Q$,	the set of rational numbers.

BACKWARD MARTINGALES

Let $\{X_n;\ n=0,\ -1,\ -2,\ \ldots\}$ be a submartingale with respect to $\{\mathscr{F}_n;\ n=0,\ -1,\ -2,\ \ldots\}$. For a concrete example, one might suppose $\mathscr{F}_n$ to be the σ-field generated by some jointly distributed random variables $\{Y_n,\ Y_{n-1}, Y_{n-2},\ \ldots\}$, but other situations are, of course, possible.

The maximal inequality, Lemma 3.1, becomes

$$\lambda \Pr\left\{\max_{n\le k\le 0} X_k > \lambda\right\} \le E[X_0], \qquad \lambda > 0, \tag{8.1}$$

provided every $X_n \ge 0$, and, in view of the independence of the right-hand side on n,

$$\lambda \Pr\left\{\sup_{k\le 0} X_k > \lambda\right\} \le E[X_0], \qquad \lambda > 0.$$

We discover the same improvement in the upcrossings inequality (4.11) of Section 4. Given real numbers $a < b$ and a negative integer N, define $V_{a,b}(N)$ to be the number of pairs (i, j), $N \le i < j \le 0$, for which the inequalities $X_i \le a$, $a < X_k < b$, for $i < k < j$, and $X_j \ge b$ take place. That is, $V_{a,b}(N)$ counts the number of times X_n upcrosses the interval (a, b) for $N \le n \le 0$, with n traversing from N to 0. Then from Eq. (4.11)

$$\begin{aligned} E[V_{a,b}(N)] &\le (b-a)^{-1}\{E[(X_0 - a)^+] - E[(X_N - a)^+]\} \\ &\le (b-a)^{-1}E[(X_0 - a)^+]. \end{aligned}$$

Again, the right-hand side does not depend on N, so that

$$E[V_{a,b}] \le (b-a)^{-1}E[(X_0 - a)^+],$$

where $V_{a,b} = V_{a,b}(-\infty)$ is the number of upcrossings of (a, b) by X_n for all $n \le 0$.

As a consequence of these strengthened inequalities, a martingale $\{X_n\}$ whose index set is $\{\ldots, -2, -1, 0\}$ *always* possesses a limit as $n \to -\infty$: needing no additional assumptions,

$$X_{-\infty} = \lim_{n\to -\infty} X_n$$

exists with probability one. But even more is true. If

$$\{X_n;\ n = 0, -1, -2, \ldots\}$$

is a martingale, $\{|X_n|\}$ is a submartingale, and by (8.1)

$$\Pr\{W > \lambda\} \to 0, \qquad \text{as} \quad \lambda \to \infty,$$

where $W = \sup|X_n|$. Using the submartingale property, in the form $E[|X_n|I(A_n)] \leq E[|X_0|I(A_n)]$ for any event A_n that is $\mathscr{F}_n$-measurable, to justify the first inequality, we get

$$\sup_{n\leq 0} E[|X_n|I\{|X_n| > c\}] \leq \sup_{n\leq 0} E[|X_0|I\{|X_n| > c\}]$$
$$\leq E[|X_0|I\{W > c\}].$$

The same reasoning as in Eq. (6.8) shows that the last term goes to zero as $c \to \infty$. Thus, the martingale $\{X_n;\ n = 0, -1, -2 \ldots\}$ is uniformly integrable, and

$$\lim_{n\to-\infty} E[|X_n - X_{-\infty}|] = 0.$$

Naturally, this reasoning applies instantly to a martingale

$$\{X_n;\ n = \ldots, -1, 0, +1, \ldots\}$$

indexed by the set of all integers. For such a martingale,

$$X_{-\infty} = \lim_{n\to-\infty} X_n$$

always exists,

$$E[|X_n - X_{-\infty}|] \to 0, \qquad \text{as} \quad n \to -\infty,$$

and, furthermore,

$$E[X_{-\infty}] = E[X_n] = E[X_0], \qquad \text{for all} \quad n.$$

In striking contrast, following the basic martingale convergence theorem, something additional, say,

$$\sup_{n\geq 0} E[X_n^+] < \infty,$$

is essential in order to secure the existence of

$$X_{+\infty} = \lim_{n\to+\infty} X_n,$$

and the equation

$$E[X_{+\infty}] = E[X_0]$$

requires even more hypotheses, e.g., that the sequence $\{X_n;\ n \geq 0\}$ be uniformly integrable.

Let $\{Z_n;\ n = 0, 1, \ldots\}$ be random variables on a probability space $(\Omega, \mathscr{F}, P)$ and let $\{\mathscr{G}_n;\ n = 0, 1, \ldots\}$ be a *decreasing* sequence of sub-σ-fields of $\mathscr{F}$, viz.,

$$\mathscr{F} \supset \mathscr{G}_n \supset \mathscr{G}_{n+1}, \qquad \text{for all} \quad n.$$

Then $\{Z_n\}$ is called a *backward martingale with respect to* $\{\mathcal{G}_n\}$ if for $n = 0, 1, \ldots$

(i) Z_n is $\mathcal{G}_n$-measurable,
(ii) $E[|Z_n|] < \infty$, and
(iii) $E[Z_n|\mathcal{G}_{n+1}] = Z_{n+1}$.

Thus $\{Z_n\}$ is a backward martingale, if and only if

$$X_n = Z_{-n}, \qquad n = 0, -1, -2, \ldots,$$

forms a martingale with respect to

$$\mathcal{F}_n = \mathcal{G}_{-n}, \qquad n = 0, -1, -2, \ldots.$$

On the basis of our preceding disucssion, the following *backward martingale convergence theorem* is established.

Theorem 8.1. *Let* $\{Z_n\}$ *be a backward martingale with respect to a decreasing sequence of* σ*-fields* $\{\mathcal{G}_n\}$. *Then with probability one*

$$Z = \lim_{n\to\infty} Z_n$$

exists,

$$\lim_{n\to\infty} E[|Z - Z_n|] = 0,$$

and

$$E[Z_n] = E[Z], \qquad \textit{for all} \quad n.$$

Example: The Law of Large Numbers. Let $X_1, X_2, \ldots$ be independent identically distributed random variables for which $E[|X_1|] < \infty$. Let $\mu = E[X_1]$, $S_0 = 0$, and introduce the partial sum $S_n = X_1 + \cdots + X_n$ for $n \geq 1$. Let $\mathcal{G}_n$ be the σ-field generated by $\{S_n, S_{n+1}, \ldots\}$. We will derive the strong law of large numbers from the observation that

$$Z_n = n^{-1}S_n \qquad (Z_0 = \mu)$$

forms a backward martingale with respect to $\mathcal{G}_n$. Clearly, $E[|Z_n|] < \infty$ and Z_n is $\mathcal{G}_n$-measurable.

We start with the trivial identity

$$\begin{aligned} S_n &= E[S_n|S_n, S_{n+1} \ldots] \\ &= \sum_{k=1}^{n} E[X_k|\mathcal{G}_n] \\ &= nE[X_k|\mathcal{G}_n], \qquad 1 \leq k \leq n, \end{aligned}$$

the last equality resulting in view of the symmetry of the summands. It is convenient to write this relation in the form

$$E[X_k|\mathcal{G}_n] = n^{-1}S_n = Z_n, \qquad 1 \leq k \leq n.$$

It follows that

$$\begin{aligned} E[Z_{n-1}|\mathcal{G}_n] &= (n-1)^{-1}E[S_{n-1}|\mathcal{G}_n] \\ &= (n-1)^{-1}\sum_{k=1}^{n-1} E[X_k|\mathcal{G}_n] \\ &= Z_n, \end{aligned}$$

which verifies the backward martingale property. [The full independence is not required in order that $\{Z_n\}$ be a backward martingale. A weaker sufficient condition is that $\{X_k\}$ be *exchangeable* (also called *symmetric* or *interchangeable*) random variables, meaning that $(X_1, \ldots, X_n)$ have the same joint distribution as $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ for every integer $n \geq 0$ and every permutation σ of the indices $(1, \ldots, n)$ into themselves.]

Invoking the backward martingale convergence theorem, we find that

$$Z = \lim_{n\to\infty} Z_n \qquad \text{exists with probability one,}$$

and $E[Z] = E[Z_n] = \mu$. The independence of $X_1, X_2, \ldots$ is vital in order to conclude that Z is nonrandom, so that, in fact, $Z \equiv \mu$. The proof follows. For any $m = 1, 2, \ldots,$

$$Z = \lim_{n\to\infty} \frac{X_m + X_{m+1} + \cdots + X_{n+m}}{n},$$

since any finite number of terms bears no influence in the limit. It follows that Z and $Z_m = m^{-1}S_m$ are independent for any finite $m = 1, 2, \ldots$. Hence, for any real a,

$$\Pr\{Z \geq a \text{ and } Z_m \geq a\} = \Pr\{Z \geq a\}\Pr\{Z_m \geq a\},$$

$$\Pr\{Z \geq a \text{ and } \max_{n\leq k\leq m} Z_k \geq a\} = \Pr\{Z \geq a\}\Pr\{\max_{n\leq k\leq m} Z_k \geq a\},$$

and

$$\Pr\{Z \geq a \text{ and } \limsup Z_n \geq a\} = \Pr\{Z \geq a\}\Pr\{\limsup Z_n \geq a\}.$$

But $Z = \lim Z_n = \limsup Z_n$, and therefore

$$\Pr\{Z \geq a\} = [\Pr\{Z \geq a\}]^2.$$

It follows that $\Pr\{Z \geq a\}$ can only attain the values 0 or 1, for every real a, and this property implies that Z is constant (why?). Moreover, in view

of $E[Z] = \mu$, the constant value of Z must be μ. We have completed the proof of the strong law of large numbers

$$\lim_{n\to\infty} n^{-1}S_n = \mu$$

with probability one.

CONTINUOUS PARAMETER MARTINGALES

Let $\{X(t);\ t \geq 0\}$ be a continuous parameter stochastic process on a probability space $(\Omega, \mathcal{F}, P)$. For each $t \geq 0$, let $\mathcal{F}_t$ be a sub-σ-field of $\mathcal{F}$ with

$$\mathcal{F}_s \subset \mathcal{F}_t, \qquad \text{if} \quad s \leq t.$$

A random variable T, having possible values in $[0, \infty]$, is called a Markov time relative to $\{\mathcal{F}_t\}$ if, for every $t \geq 0$, the event $\{T \leq t\}$ is in $\mathcal{F}_t$. We may think of $\mathcal{F}_t$ as the information available up to time t. From this viewpoint, the event that a Markov time is less than or equal to t is completely decidable by the information available up to time t.

Since a σ-field includes the complement set of each of its members, an equivalent requirement is

$$\{T > t\} \in \mathcal{F}_t, \qquad \text{for all} \quad t > 0. \tag{8.2}$$

For continuous parameter processes, it is not sufficient to require $\{T = t\}$ *to be an event in* $\mathcal{F}_t$ *for each* t. However, as before, every constant time $T \equiv \tau$ is a Markov time, and if S and T are Markov times, so are

$$S + T, \qquad S \wedge T = \min\{S, T\}, \qquad \text{and} \qquad S \vee T = \max\{S, T\}.$$

Thus, if T is a Markov time, so is $T \wedge t = \min\{T, t\}$ for every fixed $t > 0$.

Of fundamental importance are the times T_a where the process values first reach a given level a or beyond,

$$T_a = \inf\{t \geq 0;\ X(t) \geq a\}.$$

Let us suppose that every path $X(t)$ is a continuous function of t. This will be the case, for example, if $X(t)$ is Brownian motion. Let $\mathcal{F}_t = \mathcal{F}(X(s);\ 0 \leq s \leq t)$ be the σ-field generated by $\{X(s);\ 0 \leq s \leq t\}$. Then each $X(s)$ is $\mathcal{F}_s$-measurable and $\mathcal{F}_s \subset \mathcal{F}_t$ for $s < t$. In this context, T_a is a Markov time with respect to $\{\mathcal{F}_t\}$. To verify (8.2), observe that, $X(t)$ being continuous, $\{T > t\}$ is synonymous with the occurrence, for some $k = 1, 2 \ldots$, of the event $\{\min_{0 \leq u \leq t}(a - X(u)) \geq 1/k\}$, which, again using the continuity, is equivalent to the simultaneous occurrence of $\{(a - X(r)) \geq 1/k\}$ for every rational r, $0 \leq r \leq t$. That is,

$$\{T > t\} = \bigcup_{k=1}^{\infty} \bigcap_{\substack{r,\ \text{rational} \\ 0 \leq r \leq t}} \{(a - X(r)) \geq 1/k\}.$$

Each event $\{(a - X(r)) \geq 1/k\} \in \mathscr{F}_t$ as $r \leq t$. Since $\mathscr{F}_t$ is a σ-field, and there are only denumerably many rationals r in $[0, t]$.

$$\bigcap_{0 \leq r \leq t} \{(a - X(r)) \geq 1/k\} \in \mathscr{F}_t .$$

Again the union of denumerably many sets in $\mathscr{F}_t$ is itself in $\mathscr{F}_t$, and thus $\{T > t\} \in \mathscr{F}_t$ as we wished to show.

Let A be a closed set and define $T(A)$, the *entry time to* A, to be the random time

$$T(A) = \inf\{t \geq 0 ; X(t) \in A\}.$$

The parallel reasoning reveals that $T(A)$ is a Markov time with respect to $\mathscr{F}_t = \mathscr{F}(X(u); 0 \leq u \leq t)$, provided $X(t)$ is a continuous function of t.

Unfortunately, for several technical reasons, the entry time to a set A is not necessarily a Markov time with respect to $\{\mathscr{F}_t\}$ if $X(t)$ is not continuous or A not closed. It is possible to repair this defect however, by enlarging the σ-fields $\mathscr{F}_t$. Suppose that every realization $X(t)$, as a function of t, is continuous from the right and possesses a limit from the left. That is, suppose

$$X(t) = \lim_{s \downarrow t} X(s), \qquad \text{for all} \quad t \geq 0,$$

and

$$X(t-) = \lim_{s \uparrow t} X(s) \qquad \text{exists for all} \quad t > 0.$$

Continuing with $\mathscr{F}_t = \mathscr{F}(X(u); 0 \leq u \leq t)$, let $\mathscr{F}_{t^+}$ consist exactly of those events that are in every σ-field $\mathscr{F}_{t+\varepsilon}$ for every $\varepsilon > 0$. In set-theoretic terms, $\mathscr{F}_{t^+}$ is the intersection

$$\mathscr{F}_{t^+} = \bigcap_{\varepsilon > 0} \mathscr{F}_{t+\varepsilon} .$$

Each $\mathscr{F}_{t_+}$ is a σ-field, each $X(t)$ is $\mathscr{F}_{t^+}$-measurable, and

$$\mathscr{F}_{s^+} \subset \mathscr{F}_{t^+}, \qquad \text{if} \quad s < t.$$

Finally, let $\overline{\mathscr{F}}_{t^+}$ be the smallest σ-field containing every set in $\mathscr{F}_{t^+}$ together with every set in Ω that is a subset of a set $A \in \mathscr{F}$ for which $P[A] = 0$. Roughly speaking, $\overline{\mathscr{F}}_{t^+}$ consists of all events that are probabilistically equivalent to events in $\mathscr{F}_{t^+}$.

Then for every Borel set A, the entry time

$$T(A) = \begin{cases} \inf\{t \geq 0 : X(t) \in A\}, & \text{if} \quad X(t) \in A \quad \text{for some } t \geq 0, \\ \infty, & \text{if} \quad X(t) \notin A \quad \text{for all} \quad t, \end{cases}$$

is a Markov time with respect to $\{\overline{\mathscr{F}}_{t^+}\}$.

Both the martingale optional sampling and convergence theorems are valid in continuous time. If $\{X(t);\ t \geq 0\}$ is a submartingale with respect to $\{\mathcal{F}_t\}$, then

$$E[X(0)] \leq E[X(T \wedge t)] \leq E[X(t)], \qquad t \geq 0, \tag{8.3}$$

for all Markov times T. The inequalities are reversed for a supermartingale and equality obtains for a martingale. If $\Pr\{T < \infty\} = 1$, then

$$X(T \wedge t) \to X(T), \qquad \text{as} \quad t \to \infty.$$

The optional sampling theorem results when we justify the interchange of this limit with the expectation in (8.3).

Theorem 8.1. *Let $\{X(t);\ t \geq 0\}$ be a submartingale and T a Markov time with respect to $\{\mathcal{F}_t\}$. If $\Pr\{T < \infty\} = 1$ and the random variables $\{X(t \wedge T)^+;\ t \geq 0\}$ are uniformly integrable, then*

$$E[X(0)] \leq E[X(T)].$$

Corollary 8.1. *Let $\{X(t);\ t \geq 0\}$ be a martingale and T a Markov time. If $\Pr\{T < \infty\} = 1$ and $E[\sup_{t \geq 0} |X(t)|] < \infty$, then*

$$E[X(0)] = E[X(T)].$$

We use these results to derive a number of important properties of Brownian motion in Chapter 7. If $\{X(t);\ t \geq 0\}$ is a Brownian motion process with mean zero and variance parameter σ^2, then all of

(i) $X(t)$,
(ii) $Y(t) = X^2(t) - \sigma^2 t$, and
(iii) $Z(t) = \exp\{\theta X(t) - \frac{1}{2}\theta^2\sigma^2 t\}$, real θ,

are martingales with respect to $\mathcal{F}_t = \mathcal{F}(X(u);\ 0 \leq u \leq t)$. This is proved in Section 5 of Chapter 7 where these martingales are used to derive a number of probabilistic quantities associated with Brownian motion.

It is also true, but much more difficult to show, that

$$W(t) = \exp\left\{\theta f[X(t)] - \frac{\sigma^2}{2}\int_0^t \{\theta^2(f'[X(u)])^2 + \theta f''[X(u)]\}\, du\right\}$$

is a martingale for every real θ and every strictly increasing function f having continuous first and second derivatives f' and f'', respectively, provided, as usual, $E[W(t)] < \infty$.

Poisson Processes

If $\{X(t);\ t \geq 0\}$ is a Poisson process with parameter λ, then all of

$$Y(t) = X(t) - \lambda t \tag{8.4}$$

$$U(t) = Y^2(t) - \lambda t, \tag{8.5}$$

and

$$V(t) = \exp\{-\theta X(t) + \lambda t(1 - e^{-\theta})\}, \qquad -\infty < \theta < \infty, \tag{8.6}$$

are martingales [with respect to $\mathcal{F}_t = \mathcal{F}(X(u);\ 0 \leq u \leq t)$].

Fix a positive integer a and let T_a be the first time $X(t)$ reaches a. Applying the optional sampling theorem to (8.4)–(8.6) under the assumption $X(0) = 0$ and with the observation $X(T_a) = a$ (since a Poisson process varies by unit jumps), we obtain $a = \lambda E[T_a]$,

$$E[(a - \lambda T_a)^2] = \lambda E[T_a] = a, \qquad \text{or} \qquad \text{variance}(T_a) = a/\lambda^2,$$

and with

$$\beta = -\lambda(1 - e^{-\theta}),$$
$$e^{\theta a} = E[\exp\{-\beta T_a\}],$$

or

$$E[\exp\{-\beta T_a\}] = \left(\frac{\lambda}{\lambda + \beta}\right)^a.$$

This last expression is the Laplace transform of the distribution of T_a. It shows, as we already knew, that T_a has a gamma distribution with parameters a and λ.

Birth Processes

Suppose $\{X(t);\ t \geq 0\}$ is a pure birth process having birth parameters $\lambda(i) \geq 0$ for $i \geq 0$. Assume, for convenience only, $X(0) = 0$. We claim that

$$Y(t) = X(t) - \int_0^t \lambda[X(u)]\, du,$$

and

$$V(t) = \exp\left\{\theta X(t) + [1 - e^{\theta}] \int_0^t \lambda[X(u)]\, du\right\}, \tag{8.7}$$

where θ is fixed, are both martingales, provided their expectations are finite. There are a number of ways to verify these assertions. One of the best is to reduce the problem to an equivalent assertion concerning Poisson processes.

Let $\tau_0, \tau_1, \ldots$ denote the times between successive births in the given process. The τ_i's are independent, and τ_k has an exponential distribution with parameter $\lambda(k)$. The variables $\sigma_k = \lambda(k)\tau_k$, $k = 0, 1, \ldots$, remain independent and in addition have a common exponential distribution with parameter one. They can serve as interoccurrence times in a standard Poisson process $\{N(t), t \geq 0\}$. The relation between $X(t)$ and $N(t)$ is illustrated in Fig. 1.

From what was stated earlier,

$$W(T) = \exp\{\theta N(T) + T(1 - e^{\theta})\}, \qquad T \geq 0,$$

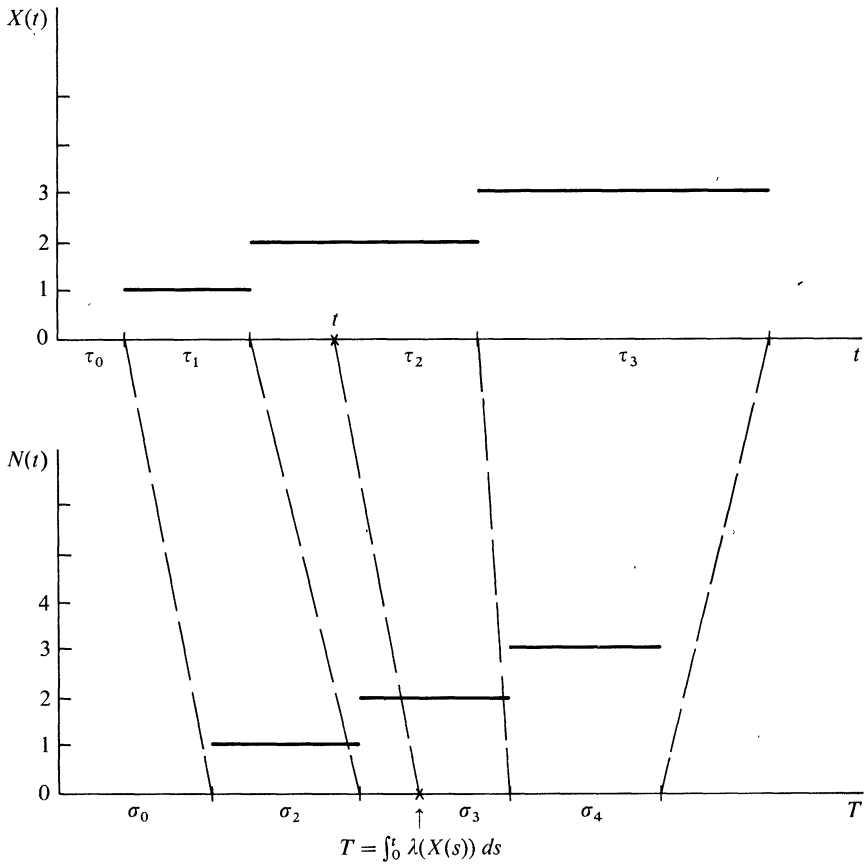


FIG. 1. The interoccurrence times $\sigma_k = \lambda(k)\tau_k$ all have mean one and thus define a Poisson process.

is a martingale with respect to $\mathcal{G}_T = \mathcal{F}(N(u); 0 \leq u \leq T)$. Thus for $T \geq S$,

$$E[W(T)|N(u); 0 \leq u \leq S] = W(S).$$

Fix a point t and let $T = T(t)$ be the corresponding point on the T scale. The relation is

$$T = T(t) = \int_0^t \lambda[X(u)]\, du; \tag{8.8}$$

and

$$N(T) = X(t). \tag{8.9}$$

Fix a point $s < t$ and let $S = S(s)$ correspond to it in a similar manner. Conditioning with respect to $\{N(u);\ 0 \leq u \leq S\}$ is equivalent to conditioning with respect to $\{X(u);\ 0 \leq u \leq s\}$. Thus

$$E[W(T)|X(u);\ 0 \leq u \leq s] = W(S). \tag{8.10}$$

But $W(T) = V(t)$ and $W(S) = V(s)$, through the substitution of (8.8) and (8.9). Thus

$$E[V(t)|\mathcal{F}_s] = V(s),$$

where $\mathcal{F}_s = \mathcal{F}(X(u);\ 0 \leq u \leq s)$ and $\{V(t)\}$ is a martingale.

The cautious reader will have noted a lacuna in our argument. For a fixed t, $T = T(t)$ is not fixed, but is random. However, T is a Markov time with respect to $\{\mathcal{G}_T\}$, and an application of an extended version of the optional sampling theorem works to verify (8.10).

Formally, we may show that $Y(t) = X(t) - \int_0^t \lambda[X(u)]\, du$ is a martingale by letting θ vanish in the martingale

$$\theta^{-1}[V(t) - 1] = Y(t) + o(\theta),$$

where $o(\theta)$ are (random) remainder terms. The left-hand side is a martingale for every $\theta \neq 0$; hence $Y(t)$ is a martingale.

Birth and Death Processes

Let $\{X(t);\ t \geq 0\}$ be a birth and death process with birth parameters $\lambda_i = \lambda(i)$, $i \geq 0$, and death parameters $\mu_i = \mu(i)$, $i \geq 1$. Assume $\lambda(0) = 0$, so that 0 is an absorbing state, but suppose $\lambda(i) > 0$ for $i \geq 1$. Define

$$f(0) = 0, \qquad f(1) = 1,$$

and

$$f(j) = 1 + \frac{\mu_1}{\lambda_1} + \frac{\mu_1\mu_2}{\lambda_1\lambda_2} + \cdots + \frac{\mu_1 \cdots \mu_{j-1}}{\lambda_1 \cdots \lambda_{j-1}}, \qquad \text{for } j \geq 2. \tag{8.11}$$

Then $Z(t) = f[X(t)]$ is a martingale whenever its mean is finite. (Compare to Elementary Problem 25). To see this, fix $s < t$ and a state $i \geq 1$, and consider

$$\begin{aligned} g_i(t) &= E[Z(t)|X(u);\ 0 \leq u \leq s,\ X(s) = i] \\ &= E[Z(t)|X(s) = i], \end{aligned}$$

the last equation resulting by the Markov property. Then for small $h > 0$, on examining the transitions that can occur over the time interval $(t, t+h)$, we obtain the equation

$$\begin{aligned} g_i(t+h) &= \sum_{k=0}^{\infty} E[Z(t+h)|X(t) = k]\ \Pr\{X(t) = k|X(s) = i\} \\ &= g_i(t) + h\sum_{k=0}^{\infty} \{\lambda_k[f(k+1) - f(k)] - \mu_k[f(k) - f(k-1)]\} \\ &\quad \cdot \Pr\{X(t) = k|X(s) = i\} + o(h). \end{aligned}$$

Thus

$$\begin{aligned} g_i'(t) &= \lim_{h \downarrow 0} \frac{g_i(t+h) - g_i(t)}{h} \\ &= \sum_{k=0}^{\infty} \{\lambda_k[f(k+1) - f(k)] - \mu_k[f(k) - f(k-1)]\} \Pr\{X(t) = k|X(s) = i\} \\ &= \sum_{k=0}^{\infty} \left\{\lambda_k\left[\frac{\mu_1 \cdots \mu_k}{\lambda_1 \cdots \lambda_k}\right] - \mu_k\left[\frac{\mu_1 \cdots \mu_{k-1}}{\lambda_1 \cdots \lambda_{k-1}}\right]\right\} \Pr\{X(t) = k|X(s) = i\} = 0. \end{aligned}$$

Since $g_i'(t) = 0$, $g_i(t) = E[Z(t)|X(s) = i]$ is a constant function of t, for $t > s$. Letting $t \downarrow s$, we conclude

$$\begin{aligned} g_i(s) &= E[Z(s)|X(s) = i] \\ &= g_i(t) = E[Z(t)|X(s) = i], \end{aligned}$$

and $Z(t)$ is a martingale.

Fix states $i < m$ and let $v(i)$ be the probability that the process is absorbed at 0 before reaching state m conditioned on $X(0) = i$. Formally,

$$T_{0,m} = \inf\{t \geq 0 : X(t) = 0 \quad \text{or} \quad X(t) = m\}.$$

We apply the optional sampling theorem to conclude that

$$f(i) = E[Z(T_{0,m})] = v(i) \cdot 0 + (1 - v(i))f(m),$$

and subsequently

$$v(i) = \frac{f(m) - f(i)}{f(m)},$$

where f is given in (8.11).

A different approach produced a similar result in Theorem 7.1 of Chapter 4.

There are numerous other martingales associated with a birth and death process. We mention two:

(a) Let $g(i)$, $i = 0, 1, \ldots$, be arbitrary, provided the expectation of

$$Y(t) = g[X(t)] - \int_0^t \{\lambda[X(u)][g(X(u)+1) - g(X(u))] - \mu[X(u)][g(X(u)) - g(X(u)-1)]\}\, du,$$

is finite. Then $\{Y(t)\}$ is a martingale.

Observe that the integral greatly simplifies when $g(i)$ is a solution to

$$\lambda(i)[g(i+1) - g(i)] - \mu(i)[g(i) - g(i-1)] \equiv 1, \qquad i \geq 1.$$

(b) Let $g(i)$, $i = 0, 1, \ldots$, be arbitrary provided the expectation of

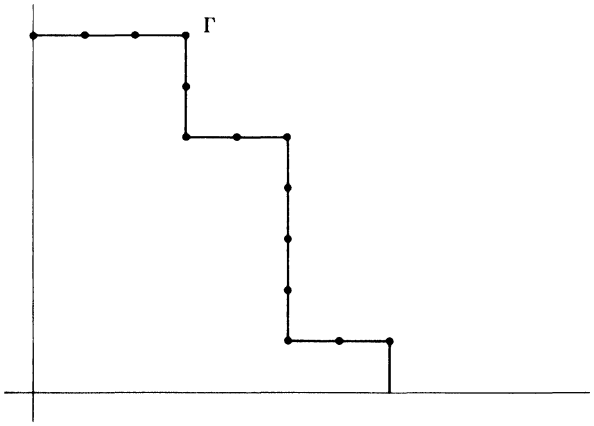
$$V(t) = \exp\left(-\theta g[X(t)] - \int_0^t [\lambda(X(u))\{1 - e^{-\theta[g(X(u)+1) - g(X(u))]}\} + \mu(X(u))\{1 - e^{+\theta[g(X(u)) - g(X(u)-1)]}\}]\, du\right)$$

is finite for some fixed real parameter θ. Then $\{V(t)\}$ is a martingale.

Elementary Problems

1. Consider a random walk on the integer lattice of the positive quadrant in two dimensions. If at any step the process is at (m, n), it moves at the next step to $(m+1, n)$ or $(m, n+1)$ with probability $\frac{1}{2}$ each. Let the process start at $(0, 0)$. Let Γ be any curve connecting neighboring lattice points (extending from the Y axis to the X axis) in the first quadrant. Show that $EY_1 = EY_2$, where Y_1 and Y_2 denote the number of steps to the right and up, respectively, before hitting the boundary Γ. The diagram describes an example of Γ.

Hint: Use an optional stopping theorem for partial sums to show $E[Y_1] = E[Y_2] = \frac{1}{2}E[T]$, where T is the number of steps it takes to reach the boundary.



2. Consider the following discrete time Markov process with the unit interval as state space. If the process is at p $(0 < p < 1)$ at the present, it will jump to $\alpha + \beta p$ with probability p and to βp with probability $1 - p$ after the next trial, where $\alpha, \beta > 0$ and $\alpha + \beta = 1$. In symbols, the process is defined by the transformation law

$$X_{n+1} = \begin{cases} \alpha + \beta X_n, & \text{with probability } X_n, \\ \beta X_n, & \text{with probability } 1 - X_n. \end{cases}$$

Show that this process is a martingale.

3. Let $W(n)$ be a branching process with immigration:

$$W(n+1) = Y_n + X_{n,1} + \cdots + X_{n,W(n)},$$

where Y_n is the immigration in generation n and $X_{n,j}$ is the number of offspring of the jth individual in generation n, all independent. Suppose $E[Y_n] = \lambda$ and $E[X_{n,j}] = m \neq 1$. Show that

$$Z_n = m^{-n}\left[W(n) - \lambda\left\{\frac{1-m^n}{1-m}\right\}\right]$$

is a martingale.

4. Let X_n be the number of males and Y_n the number of females in the nth generation of a population. Permanent pairs are formed. Thus, $Z_n = \min\{X_n, Y_n\}$ pairs produce offspring, and they do so independently according to the generating function $g(t, s) = E[t^{\xi}s^{\eta}]$, where ξ is the number of male, and η the number of female offspring of a single parental pair. Show that Z_n is a supermartingale if either $E[\xi] \leq 1$ or $E[\eta] \leq 1$ holds.

5. Assume $Y_1, Y_2, \ldots$ are independent and identically distributed with $\Pr\{Y_1 = +1\} = p$, and $\Pr\{Y_1 = -1\} = q = 1 - p$. Fix positive integers a and b. With $S_0 = 0$, and $S_n = Y_1 + \cdots + Y_n$, $n \geq 1$, let

$$T = \min\{n: \quad S_n = -a \text{ or } S_n = b\}.$$

Establish the formula

$$E[T] = \frac{b}{p-q} - \frac{a+b}{p-q} \cdot \frac{1-(p/q)^b}{1-(p/q)^{a+b}} \qquad \text{when } p \neq q.$$

(The equation $E[T] = ab$ when $p = q = \frac{1}{2}$ was derived in Example (a) of Section 4.)

6. Let $Y_1, Y_2, \ldots$ be the independent and identically distributed with $\Pr\{Y_1 = +1\} = p$, and $\Pr\{Y_1 = -1\} = q = 1 - p$. Suppose $p > \frac{1}{2} > q$. With $S_0 = 0$ and $S_n = Y_1 + \cdots + Y_n$ for $n \geq 1$ let

$$T = \min\{n: \quad S_n \geq b\}$$

for some fixed positive integer b. Deduce the generating function

$$E[s^T] = \left(\frac{1-\{1-4pqs^2\}^{1/2}}{2qs}\right)^b, \qquad 0 < s \leq 1,$$

Hint: Use Wald's identity.

7. Under the conditions of Elementary Problem 6, derive the mean $E[T] = b/(p-q)$ and the variance $\mathrm{Var}[T] = b[1-(p-q)^2]/(p-q)^3$.

8. Let $Y(0)$, $Y(1)$, ... be the success-runs Markov chain in which $P_{00} = 1$, so 0 is an absorbing state, and $P_{i,i+1} = p$, $P_{i0} = q = 1-p$ for $i = 1, 2, \ldots$. Let a and b be arbitrary real constants. Show that

$$X_n = \begin{cases} b & \text{if } Y(n) = 0, \\ a\left(\frac{1}{p}\right)^{Y(n)-1} + b\left[1 - \left(\frac{1}{p}\right)^{Y(n)-1}\right], & \text{if } Y(n) > 0 \end{cases}$$

is a martingale.

9. Consider a family of r.v.'s $\{X_n\}_0^\infty$, each having finite absolute expectation and satisfying

$$E[X_{n+1}|X_0, X_1, \ldots, X_n] = \alpha X_n + \beta X_{n-1}, \qquad n > 0,$$

with $\alpha > 0$, $\beta > 0$, and $\alpha + \beta = 1$. Find an appropriate value of a such that the sequence

$$Y_n = aX_n + X_{n-1}, \qquad n \geq 1, \quad Y_0 = X_0,$$

constitutes a martingale with respect to $\{X_n\}$.

10. Let $\{X_n\,;\, n \geq 0\}$ be a martingale with respect to $\{Y_n\}$. Prove for any set of integers $k \leq l < m$ that the difference $X_m - X_l$ is uncorrelated with X_k, that is,

$$E[(X_m - X_l)X_k] = 0.$$

Hint: Evaluate the expectation by conditioning on $Y_1, \ldots, Y_k$.

11. Let $\{\xi_i\}$ be a sequence of r.v.'s such that the partial sums

$$X_n = \xi_0 + \xi_1 + \cdots + \xi_n, \qquad n \geq 1,$$

determine a martingale. Show that the summands are mutually uncorrelated, i.e., $E[\xi_i \xi_j] = 0$ for $i \neq j$.

12. Let $S_n = X_1 + \cdots + X_n$ be a martingale satisfying $E[X_k^2] \leq K < \infty$, for all k. Show that S_n obeys the weak law of large numbers:

$$\Pr\{|S_n/n| > \varepsilon\} \to 0, \qquad \text{as } n \to \infty,$$

for any positive ε.

Hint: Use the maximal inequality and the orthogonality result of Elementary Problem 11.

13. Let $\{\xi_i\}_{i=0}^{\infty}$ be a sequence of real valued jointly distributed random variables that satisfy $E[\xi_i|\xi_0, \xi_1, \ldots, \xi_{i-1}] = 0$, $i = 1, 2, \ldots$. Define

$$X_0 = \xi_0, \qquad X_{n+1} = \sum_{i=0}^{n} \xi_{i+1} f_i(\xi_0, \xi_1, \ldots, \xi_i),$$

where f_i are a prescribed sequence of functions of $i + 1$ real variables. Show that $\{X_n\}$ form a martingale.

14. Consider a game of tossing repeatedly and independently a fair coin where the result ξ_k at round k has $\Pr\{\xi_k = 1\} = \Pr\{\xi_k = -1\} = \frac{1}{2}$. Suppose a player stakes in the first round a unit and doubles the stake each time he loses and returns to the unit stake each time he wins. Assume the player has unlimited funds (or credit). Let X_n be the net gain after the nth round. Show that $\{X_n\}_1^{\infty}$ determines a martingale with respect to $\{\xi_n\}_1^{\infty}$.

Hint: Establish that X_n can be represented in the form

$$X_n = \sum_{k=1}^{n} \xi_k f_k(\xi_1, \ldots, \xi_{k-1})$$

for suitable f_k, and consult Elementary Problem 13.

15. (a) Consider a Markov chain $\{X_n;\ n > 0\}$ on the state space $\{0, 1, 2, \ldots, N\}$ with transition probability matrix

$$(*) \qquad P_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

Establish that $\{X_n;\ n \geq 0\}$ and $\left\{V_n = \dfrac{X_n(N - X_n)}{(1 - N^{-1})^n},\ n \geq 0\right\}$ constitute martingales with respect to $\{X_n\}$.

(b) Replace (*) by

$$(**) \qquad P_{ij} = \frac{\binom{2i}{j}\binom{2N-2i}{N-j}}{\binom{2i}{N}}.$$

In this case determine λ such that

$$W_n = \frac{X_n(N - X_n)}{\lambda^n}, \qquad n \geq 0,$$

is a martingale with respect to $\{X_n\}$.

16. Suppose Y_0 is uniformly distributed on $(0, 1]$, and given Y_n, suppose Y_{n+1} is uniformly distributed on $(1 - Y_n, 1]$. Show $X_0 = Y_0$, and

$$X_n = 2^n \prod_{k=1}^{n} \left[\frac{1 - Y_k}{Y_{k-1}}\right], \qquad n = 1, 2, \ldots$$

is a martingale.

17. From an urn that initially contains one red and one green ball, a ball is drawn at random and it and one more of the same color are returned. This process is repeated indefinitely. Let X_n be the fraction of red balls at stage n. (a) Use the maximal inequality to show $\Pr\{X_n \geq 3/4 \text{ for some } n = 1, 2, \ldots\} \leq 2/3$. In words, there is 2/3 or less chance of ever there being more than 3/4 of the balls red. (b) Using the limit distribution found in Example (g) of Section 6, show $\lim_{n\to\infty} \Pr\{X_n \geq 3/4\} = 1/4$. In words, the probability is 1/4 that, in the limit, 3/4 or more of the balls will be red.

18. Let $P_{ij} = e^{-i} i^j/j!$, $i, j = 0, 1, \ldots$ be the transition probabilities for a Markov chain X_n. We consider $P_{00} = 1$. (a) Verify that X_n is a martingale. (b) Derive the inequality

$$\Pr\{\max_{0 \leq n < \infty} X_n \geq a | X_0 = i\} \leq i/a$$

for $i, a = 1, 2, \ldots$. (c) Prove that $\lim_{n\to\infty} X_n = 0$ with probability one.

Hint: Apply the optional sampling theorem with T being the first time n that $X_n = 0$ or $X_n \geq a$.

19. Let X_n be a Markov chain whose transition probabilities are $P_{i,j} = 1/[e(j-i)!]$ for $i = 0, 1, \ldots$ and $j = i, i+1, \ldots$.

Verify the martingale property for:

(a) $Y_n = X_n - n$;

(b) $U_n = Y_n^2 - n$;

(c) $V_n = \exp\{X_n - n(e-1)\}$.

20. Let $\{X_n\}_1^\infty$ be a submartingale. Show that the sequence

$$U_1 = 0, \qquad U_n = \sum_{i=2}^{n} \{E[X_i | X_1, X_2, X_3, \ldots, X_{i-1}] - X_{i-1}\}, \qquad n \geq 2,$$

is an increasing process, i.e., $U_n \geq U_{n-1}$.

21. Consider a Markov chain $\{X_n;\ n \geq 0\}$ on the state space of the non-negative integers with transition probability matrix $P = \|P_{ij}\|$. Let $u(i, n)$ be a function defined on the integers i, $n \geq 0$ and satisfying the functional equation

$$u(i, n) = \sum_{k=0}^{\infty} u(k, n+m) P_{ik}^{(m)}$$

where $P_{ik}^{(m)}$ is the m step transition probability from state i to k. Show that

$$U_n = u(X_n, n)$$

is a martingale with respect to $\{X_n\}$.

22. Consider a Markov chain $\{X_n;\ n \geq 0\}$ involving N states whose possible state values are $x_0 < x_1 < \cdots < x_N$ with transition possibility matrix $P_{ij} = \Pr\{X_{n+1} = x_j | X_n = x_i\}$. Suppose $\{X_n\}$ is also a martingale. Show that states x_0 and x_N are absorbing, i.e., $P_{0,0} = P_{N,N} = 1$.

23. Consider a Markov chain $\{X_n;\ n \geq 0\}$ of N states $\{0, 1, 2, \ldots N\}$ and transition probabilities

$$P_{ij} = \binom{N}{j} \pi_i^j (1 - \pi_i)^{N-j},$$

where

$$\pi_i = \frac{1 - e^{-2ai/N}}{1 - e^{-2a}}.$$

Show that $Z_n = e^{-2aX_n}$ is a Martingale.

24. Let $\{X_n,\ n \geq 0\}$ describe a transient Markov chain on the non-negative integers with transition probability matrix $P = \| P_{ij} \|$. Define

$$u(i) = \sum_{n=0}^{\infty} P_{i0}^{(n)}.$$

Show that

$$U_n = u(X_n) \text{ is a submartingale.}$$

25. Consider a finite birth and death process $\{X(t),\ t \geq 0\}$ with infinitesimal parameters λ_i and μ_i, $0 \leq i \leq N$ $(\mu_0 = 0)$. The infinitesimal matrix is

$$A = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0 \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \lambda_N & -\lambda_N \end{pmatrix}$$

Consider any solution $y = (y_0, y_1, \ldots)$ of the linear system

$$Ay = 0.$$

Establish that $Y(t) = y_{X(t)}$, $t > 0$, is a martingale with respect to $\mathcal{F}_t = \sigma(X(u);\ u \leq t)$.

Hint: Show that if $Ay = 0$, then

$$y_i = \sum_{j=0}^{N} P_{ij}(t) y_j, \qquad i = 0, 1, \ldots, N,$$

holds for all $t > 0$, where $P_{ij}(t)$ represents the transition probability matrix of the process $\{X(t),\ t \geq 0\}$.

Problems

1. Prove: if $\{X_n\}$ is a submartingale and $\varphi(x)$ is a convex, increasing function, then $\{\varphi(X_n)\}$ is a submartingale whenever $E|\varphi^+(X_n)| < \infty$ for all n (cf. Lemma 2.2).

2. Suppose $P = \|P_{ij}\|$ is the transition probability matrix of an irreducible recurrent Markov chain $\{X_n\}$. Use the supermartingale convergence theorem (see Remark 5.1) to show that every nonnegative solution $y = \{y(i)\}$ to the system of inequalities

$$y(i) \geq \sum_{j=0}^{\infty} P_{ij} y(j), \qquad \text{for all} \quad i,$$

is constant.

Hint: Paraphrase Example (a) of Section 6.

3. Let $\{U_n\}$ and $\{V_n\}$ be martingales with respect to the same process $\{Y_n\}$. Suppose $U_0 = V_0 = 0$ and $E[U_n^2] < \infty$, $E[V_n^2] < \infty$ for all n. Show

$$E[U_n V_n] = \sum_{k=1}^{n} E[(U_k - U_{k-1})(V_k - V_{k-1})].$$

As a special case,

$$E[U_n^2] = \sum_{k=1}^{n} E[(U_k - U_{k-1})^2].$$

Hint: Because $U_n V_n = \sum_{k=1}^{n} (U_k V_k - U_{k-1} V_{k-1})$, it is enough to show $E[U_k V_k - U_{k-1} V_{k-1}] = E[(U_k - U_{k-1})(V_k - V_{k-1})]$. But $E[(U_k - U_{k-1})(V_k - V_{k-1})] = E[U_k V_k] - E[U_{k-1} V_k] - E[(U_k - U_{k-1}) V_{k-1}]$. Now evaluate the last two expectations by first conditioning on $Y_0, \ldots, Y_{k-1}$ and using the martingale property.

4. Suppose $\{X_n\}$ is a martingale satisfying, for some $\alpha > 1$,

$$E[|X_n|^\alpha] < \infty, \qquad \text{for all} \quad n.$$

Show

$$E\left[\max_{0 \leq k \leq n} |X_k|\right] \leq \frac{\alpha}{\alpha - 1} E[|X_n|^\alpha]^{1/\alpha}.$$

Hint: $E[\max_{0 \leq k \leq n} |X_k|] = \int_0^\infty \Pr\{\max_{0 \leq k \leq n} |X_k| > t\}\, dt$. Now use the maximal inequality on the submartingale $|X_n|^\alpha$.

5. Let $\{X_n\}$ be a submartingale. Strengthen the maximal inequality, Lemma 5.1., to

$$\lambda \Pr\left\{\max_{0 \leq k \leq n} X_k > \lambda\right\} \leq E\left[X_n I\left\{\max_{0 \leq k \leq n} X_k > \lambda\right\}\right]$$
$$\leq E[X_n^+] \leq E[|X_n|], \qquad \lambda > 0.$$

(*Note:* Lemma 5.1 requires $X_k \geq 0$ for all k. The above does not.) Consequently, for a martingale $\{X_n\}$,

$$\lambda \Pr\left\{\max_{0 \leq k \leq n} |X_k| > \lambda\right\} \leq E\left[|X_n| I\left\{\max_{0 \leq k \leq n} |X_k| > \lambda\right\}\right], \qquad \lambda > 0.$$

6. The result of Problem 5 can be used to strengthen the inequality in Problem 4 to the form

$$E\left[\max_{0\le k\le n}|X_k|^\alpha\right]\le\left(\frac{\alpha}{\alpha-1}\right)^\alpha E[|X_n|^\alpha].$$

Prove this when $\alpha=2$.

7. *Extinction of populations.* Consider a population of organisms living in some bounded environment, say the Earth. Let X_n be the number of organisms alive at time n and observe that $\{0\}$ is an absorbing state, $X_n=0$ implies $X_{n+m}=0$ for all m. It is reasonable to suppose that for every N there exists $\delta>0$ satisfying

$$\Pr[X_{n+1}=0|X_1,\ldots,X_n]\ge\delta,\qquad\text{if}\quad X_n\le N,$$

$n=1,2,\ldots$. Let $\mathscr{E}$ be the event of eventual extinction

$$\mathscr{E}=\{X_k=0\text{ for some }k=1,2,\ldots\}.$$

Show that with probability one, either $\mathscr{E}$ occurs or else $X_n\to\infty$ as $n\to\infty$. Since the latter cannot occur in a bounded environment, eventual extinction is certain.

8. Let Z, Y_0, Y_1, ... be jointly distributed random variables and assume $E[|Z|^2]<\infty$. Show that $X_n=E[Z|Y_0,\ldots,Y_n]$ satisfies the conditions for the martingale mean square convergence theorem.

9. Let $\{X_n\}$ be a martingale satisfying $E[X_n^2]\le K<\infty$ for all n. Suppose

$$\lim_{n\to\infty}\sup_{m\ge1}|E[X_nX_{n+m}]-E[X_n]E[X_{n+m}]|=0.$$

Show that $X=\lim_{n\to\infty}X_n$ is a constant, i.e., nonrandom.

10. Let $\{X_n\}$ be a martingale for which $E[X_n]=0$ and $E[X_n^2]<\infty$ for all n. Show that

$$\Pr\left\{\max_{0\le k\le n}X_k>\lambda\right\}\le\frac{E[X_n^2]}{E[X_n^2]+\lambda^2},\qquad\lambda>0.$$

Hint: For every $c>0$, $(X_n+c)^2$ is a submartingale, and for $\lambda>0$ we may apply the maximal inequality to get

$$\begin{aligned}\Pr\left\{\max_{0\le k\le n}X_k>\lambda\right\}&\le\Pr\left\{\max_{0\le k\le n}(X_k+c)^2>(\lambda+c)^2\right\}\\&\le\frac{E[(X_n+c)^2]}{(\lambda+c)^2},\qquad\text{for all }c>0.\end{aligned}$$

Now determine the value c which gives the best bound, i.e., minimizes the right-hand side.

11. Let $\{X_n\}$ be a submartingale. Show that

$$\lambda \Pr\left\{\min_{0\le k\le n} X_k < -\lambda\right\} \le E[X_n^+] - E[X_0], \qquad \lambda > 0.$$

12. Prove: If $\{X_n\}$ is a nonnegative supermartingale, then

$$\lambda \Pr\left\{\max_{0\le k\le n} X_k \ge \lambda\right\} \le E[X_0], \qquad \lambda > 0.$$

(Cf. Lemma 5.2.)

Problems 13–16 all occur in the same context. Let $\mathscr{B}_0 = \{B_1, B_2, \ldots\}$ be a denumerable partition of a set Ω. That is, $\Omega = \bigcup_{n=1}^{\infty} B_n$ and $B_i \cap B_j = \varnothing$ if $i \ne j$. Let $\mathscr{B}$ be the σ-field consisting of $\varnothing$, Ω and all sets that are unions of sets in $\mathscr{B}_0$, i.e., of the form

$$B = \bigcup_{k=1}^{j} B_{n(k)}, \qquad 1 \le j \le \infty, \quad \text{with} \quad B_{n(k)} \in \mathscr{B}_0.$$

13. Suppose $\mathscr{B}$ is the σ-field generated by some random variable Y (having, then, at most a denumerable number of possible values). Show that a random variable X is $\mathscr{B}$-measurable if and only if $X = f(Y)$ for some real-valued function f.

14. Suppose X_1 and X_2 are $\mathscr{B}$-measurable random variables. Show that $a_1X_1 + a_2X_2$ is $\mathscr{B}$-measurable for all real a_1, a_2.

15. Suppose Y is $\mathscr{B}$-measurable, and $E[|Y|] < \infty$. Show that $E[YZ] \ge 0$ for all bounded nonnegative $\mathscr{B}$-measurable random variables Z implies $P[\{\omega\colon Y(\omega) \ge 0\}] = 1$.

16. Show that X is $\mathscr{B}$-measurable if and only if

$$X(\omega) = \sum_{k=1}^{\infty} \alpha_k I_{B_k}(\omega),$$

for some real sequence $\{\alpha_k\}$, where

$$I_{B_j}(\omega) = \begin{cases} 1, & \text{if } \omega \in B_j, \\ 0, & \text{if } \omega \notin B_j. \end{cases}$$

In particular, observe that $X(\omega)$ is constant on each of the sets B_j.

17. Fix $\lambda > 0$. Suppose $X_1, X_2, \ldots$ are jointly distributed random variables whose joint distributions satisfy

$$E[\exp\{\lambda X_{n+1}\} | X_1, \ldots, X_n] \le 1, \qquad \text{for all } n.$$

Let $S_n = X_1 + \cdots + X_n$ ($S_0 = 0$). Establish

$$\Pr\left\{\sup_{n\ge 0} (x + S_n) > l\right\} \le e^{-\lambda(l-x)}, \qquad \text{for } x \le l.$$

Hint: Use an optional sampling theorem on the nonnegative supermartingale $\exp\{-\lambda(l - x - S_n)\}$.

18. Let X be a random variable for which

$$\Pr\{-\varepsilon \leq X \leq +\varepsilon\} = 1, \tag{A}$$

and

$$E[X] \leq -\rho\varepsilon, \tag{B}$$

where $\varepsilon > 0$ and $\rho > 0$ are given. Show that

$$E[e^{\lambda X}] \leq 1,$$

for $\lambda = \varepsilon^{-1} \log[(1+\rho)/(1-\rho)]$. Apply the result of Problem 17 to bound

$$\Pr\left\{\sup_{n \geq 0} (x + S_n) > l\right\}, \qquad \text{for} \quad x < l,$$

where $S_n = X_1 + \cdots + X_n$, and the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ satisfies (A) and (B).

19. Let X be a random variable satisfying

(a) $E[X] \leq m < 0$, and
(b) $\Pr\{-1 \leq X \leq +1\} = 1$.

Suppose $X_1, X_2, \ldots$ are jointly distributed random variables for which the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ always satisfies (a) and (b). Let $S_n = X_1 + \cdots + X_n$ ($S_0 = 0$) and for $a < x$ let

$$T_a = \min\{n: x + S_n \leq a\}.$$

Establish the inequality

$$E[T_a] \leq (1 + x - a)/|m|, \qquad a < x.$$

20. Let T, Y_0, $Y_1, \ldots$ be random variables. Suppose the possible values for T are $\{0, 1, \ldots\}$ and, for every $n \geq 0$, the event $\{T \geq n\}$ is determined by $(Y_0, \ldots, Y_n)$. Is T necessarily a Markov time with respect to $\{Y_n\}$? Provide a proof or counterexample to support your claim.

21. Let $S_n = \xi_1 + \cdots + \xi_n$, where $\{\xi_k\}$ are independent identically distributed positive random variables ($\Pr\{\xi_k > 0\} = 1$). Prove that

$$\sup_{n \geq 1} E\left[\frac{n}{a + S_n}\right] < \infty, \qquad \text{for any} \quad a > 0.$$

[The case where ξ_k assumes only integer values was treated in Example (e) of Section 6].

22. Let $Y_1, Y_2, \ldots$ be independent random variables with $\Pr\{Y_k = +1\}$ $\Pr\{Y_k = -1\} = 1/2$. Put $S_k = Y_1 + \cdots + Y_k$. Show that

$$\Pr\{S_k < k \quad \text{for all } k = 1, \ldots, N | S_N = a\} = 1 - \frac{a}{N}.$$

23. Let ξ_n be nonnegative random variables satisfying

$$E[\xi_{n+1}|\xi_1, \ldots, \xi_n] \leq \delta_n + \xi_n,$$

where $\delta_n \geq 0$ are constants and $\Delta = \sum_{n=1}^{\infty} \delta_n < \infty$. Show that with probability one, ξ_n converges to a finite random variable ξ as $n \to \infty$.

24. The Haar functions on [0, 1) are defined by

$$H_1(t) \equiv 1,$$

$$H_2(t) = \begin{cases} 1, & 0 \leq t < \frac{1}{2}, \\ -1, & \frac{1}{2} \leq t < 1, \end{cases}$$

$$H_{2^n+1}(t) = \begin{cases} 2^{n/2}, & 0 \leq t < 2^{-(n+1)}, \\ -2^{n/2} & 2^{-(n+1)} \leq t < 2^{-n}, \\ 0, & \text{otherwise}, \end{cases} \quad n = 1, 2, \ldots,$$

$$H_{2^n+j}(t) = H_{2^n+1}\left(t - \frac{j-1}{2^n}\right). \qquad j = 1, \ldots, 2^n.$$

It helps to plot the first five.

Let $f(z)$ be an arbitrary function on [0, 1] but satisfying

$$\int_0^1 |f(z)|\, dz < \infty.$$

Define $a_k = \int_0^1 f(t) H_k(t)\, dt$. Let Z be uniformly distributed on [0, 1]. Show that

$$f(Z) = \lim_{n \to \infty} \sum_{k=1}^{n} a_k H_k(Z) \qquad \text{with probability one,}$$

and

$$\lim_{n \to \infty} \int_0^1 \left| f(t) - \sum_{k=1}^{n} a_k H_k(t) \right| dt = 0.$$

25. Suppose $X_1, X_2, \ldots$ are independent random variables having finite moment generating functions $\varphi_k(t) = E[\exp\{tX_k\}]$. Show, if $\Phi_n(t_0) = \prod_{k=1}^{n} \varphi_k(t_0) \to \Phi(t_0)$ as $n \to \infty$, $t_0 \neq 0$ and $0 < \Phi(t_0) < \infty$, then $S_n = X_1 + \cdots + X_n$ converges with probability one.

26. Let 0 be an absorbing state in a success runs Markov chain $\{X_n\}$ having transition probabilities $P_{00} = 1$ and $P_{i,\, i+1} = p_i = 1 - P_{i,\, 0}$ for $i = 1, 2, \ldots$. Suppose $p_i \geq p_{i+1} \geq \ldots$, and let a be the unique value for which $ap_{a-1}/(a-1) > 1 \cdot (a+1)p_a/a$. Define

$$f(i) = \begin{cases} 0, & \text{for } i = 0, \\ ap_i p_{i+1} \cdots p_{a-1}, & \text{for } 1 \leq i < a, \\ i, & \text{for } i \geq a. \end{cases}$$

(a) Show that $f(i) \geq i$ for all $i = 0, 1, \ldots$.

(b) Show that $f(i) \geq E[f(X_{n+1})|X_n = i]$ for all i, so that $\{f(X_n)\}$ is a nonnegative supermartingale.

(c) Use (a) and (b) to verify that $f(i) \geq E[X_T|X_0 = i]$ for all Markov times T.

(d) Prove $f(i) = E[X_{T^*}|X_0 = i]$, where $T^* = \min\{n \geq 0: X_n \geq a \text{ or } X_n = 0\}$.

Thus, T^* maximizes $E[X_T|X_0 = i]$ over all Markov times T.

27. Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be a countable set and $\mathcal{F}$ the σ-field of all subsets of Ω. For a fixed N, let $X_0, X_1, \ldots, X_N$ be random variables defined on Ω and let T be a Markov time with respect to $\{X_n\}$ satisfying $0 \leq T \leq N$. Let $\mathcal{F}_n$ be the σ-field generated by $X_0, X_1, \ldots, X_n$ and define $\mathcal{F}_T$ to be the collection of sets A in $\mathcal{F}$ for which $A \cap \{T = n\}$ is in $\mathcal{F}_n$ for $n = 0, \ldots, N$. That is,

$$\mathcal{F}_T = \{A: A \in \mathcal{F} \quad \text{and} \quad A \cap \{T = n\} \in \mathcal{F}_n, \quad n = 0, \ldots, N\}.$$

Show:

(a) $\mathcal{F}_T$ is a σ-field,

(b) T is measurable with respect to $\mathcal{F}_T$,

(c) $\mathcal{F}_T$ is the σ-field generated by $\{X_0, \ldots, X_T\}$, where $\{X_0, \ldots, X_T\}$ is considered to be a variable-dimensional vector-valued function defined on Ω.

28. Suppose $S_n = X_1 + \cdots + X_n$ is a zero-mean martingale for which $E[X_n^2] < \infty$ for all n. Show that $S_n/b_n \to 0$ with probability one for any monotonic real sequence $b_1 \leq \cdots \leq b_n \leq b_{n+1} \uparrow \infty$, provided $\sum_{n=1}^{\infty} E[X_n^2]/b_n^2 < \infty$.

29. Let X_n be the total assets of an insurance company at the end of year n. In each year, n, premiums totaling $b > 0$ are received, and claims A_n are paid, so $X_{n+1} = X_n + b - A_n$. Assume $A_1, A_2, \ldots$ are independent random variables, each normally distributed with mean $\mu < b$ and variance σ^2. The company is ruined if its assets ever drop to zero or less. Show

$$\Pr\{\text{ruin}\} \leq \exp\{-2(b - \mu)X_0/\sigma^2\}.$$

30. Let $Y_1, Y_2, \ldots$ be independent identically distributed positive random variables having finite mean μ. For fixed $0 < \beta < 1$, let a be the smallest value u for which $u \geq \beta E[u \vee Y_1] = \beta E[\max\{u, Y_1\}]$. Set $f(x) = a \vee x$. Show that $\{\beta^n f(M_n)\}$ is a nonegative supermartingale, where $M_n = \max\{Y_1, \ldots, Y_n\}$ whence $a = f(0) \geq E[\beta^T f(M_T)]$ for all Markov times T. Finally establish that $a = E[\beta^{T^*} M_{T^*}]$ for $T^* = \min\{n \geq 1: Y_n \geq a\}$. Thus, T^* maximizes $E[\beta^T M_T]$ over all Markov times T.

31. Let $X, X_1, X_2, \ldots$ be independent identically distributed random variables having negative mean μ and finite variance σ^2. With $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$, set $M = \max_{n \geq 0} S_n$. In view of $\mu < 0$, we know that $M < \infty$. Assume $E[M] < \infty$. (In fact, it can be shown that this is a consequence of $\sigma^2 < \infty$.) Define $r(x) = x^+ = \max\{x, 0\}$ and $f(x) = E[(x + M - E[M])^+]$.

(a) Show $f(x) \geq r(x)$ for all x.

(b) Show $f(x) \geq E[f(x+X)]$ for all x, so that $\{f(x+S_n)\}$ is a nonnegative supermartingale [*Hint*: Verify and use the fact that M and $(X+M)^+$ have the same distribution.]

(c) Use (a) and (b) to show $f(x) \geq E[(x+S_T)^+]$ for all Markov times T. $[(x+S_\infty)^+ = \lim_{n\to\infty}(X+S_n)^+ = 0.]$

32. (Continuation). Let T^* be the Markov time

$$T^* = \begin{cases} \min\{n \geq 0: x+S_n \geq E[M]\}, & \text{if } x+S_n \geq E[M] \text{ for some } n, \\ \infty, & \text{if } x+S_n < E[M], \text{ for all } n. \end{cases}$$

Show that $f(x)=E[(x+S_{T^*})^+]$, so that T^* maximizes $E[r(S_T)]$ over all Markov times T.

33. Let $\{X_n\}$ be a success runs Markov chain having transition probabilities $P_{i,\,i+1} = p_i = 1 - P_{i,0}$, for $i = 0, 1, \ldots$. Suppose $0 < p_i < 1$ and $p_i \geq p_{i+1} \geq \cdots$. Fix $0 < \beta < 1$, and let a be the unique value for which $a\beta p_{a-1}/(a-1) > 1 \geq (a+1)\beta p_a/a$. Define

$$f(i) = \begin{cases} a\beta^{a-i} p_i \cdot p_{i+1} \cdots p_{a-1}, & \text{for } i < a, \\ i, & \text{for } i \geq a. \end{cases}$$

(a) Show that $f(i) \geq i$ for all i.

(b) Show that $f(i) \geq \beta E[f(X_n)|X_{n-1} = i]$, so that $\{\beta^n f(X_n)\}$ is a nonnegative supermartingale.

(c) Use (a) and (b) to verify that $f(i) \geq E[\beta^T X_T | X_0 = i]$ for all Markov times T.

(d) Finally, prove $f(i) = E[\beta^{T^*} X_{T^*} | X_0 = i]$, where $T^* = \min\{n \geq 0: X_n \geq a\}$. Thus, T^* maximizes $E[\beta^T X_T | X_0 = i]$ over all Markov times T.

34. Let Z_n be a Markov chain having transition matrix $P(i,j)$. Let $f(i)$ be a bounded function and define $F(i) = \sum_j P(i,j)f(j) - f(i)$ for all i. Show that

$$\frac{F(Z_1) + \cdots + F(Z_n)}{n} \to 0, \qquad \text{as} \quad n \to \infty,$$

with probability one.

Hint: Use the results of Problem 28.

35. Let $\{X_n\}$ be a martingale satisfying $\sup_n E[|X_n|] < \infty$. Derive the representation $X_n = X_n^{(1)} - X_n^{(2)}$, where $\{X_n^{(i)}\}$ are nonnegative martingales having bounded means.

Hint: $Z_n^N = E[|X_{N+1}| \mid Y_0, \ldots, Y_n]$ is increasing in N, so $Z_n = \lim_{N\to\infty} Z_n^N$ exists, is nonnegative, and $E[|Z_n|] \leq \sup_n E[|X_n|] < \infty$. Prove that $\{Z_n\}$ is a martingale, and then use $X_n^{(1)} = Z_n$ and $X_n^{(2)} = Z_n - X_n$.

36. Let $\{X_n\}$ be a submartingale having a finite mean and for which $X_0 = 0$. Derive the representation $X_n = X_n' + X_n''$, where $\{X_n'\}$ is a martingale and $X_n'' \leq X_{n+1}''$ is a nondecreasing process.

Hint: See Elementary Problem 20.

37. Let $\{X_n\}$ be a martingale for which $Y = \sup_n |X_{n+1} - X_n|$ has a finite mean. Let A_1 be the event that $\{X_n\}$ converges and A_2 the event that $\limsup X_n = +\infty$ and $\liminf X_n = -\infty$. Show that $\Pr\{A_1\} + \Pr\{A_2\} = 1$. In words, $\{X_n\}$ either converges, or oscillates very greatly indeed.

Hint: For every k, $\tilde{X}_n = X_{T \wedge n}$ converges, where $T = \min\{n: X_n \geq k\}$, because $\tilde{X}_n \leq k + Y$, so $\sup_n E[\tilde{X}_n] < \infty$. Thus the alternative to $\{X_n\}$ converging is included in the event $\limsup X_n > k$ for every k. The same analysis applies to $\{-X_n\}$.

38. Let $\varphi(\xi)$ be a symmetric function, nondecreasing in $|\xi|$, with $\varphi(0) = 0$, and such that $\{\varphi(Y_j)\}_{j=0}^n$ is a submartingale. Fix $0 = u_0 \leq u_1 \leq \cdots \leq u_n$. Show that

$$\Pr\{|Y_j| \leq u_j;\ 1 \leq j \leq n\} \geq 1 - \sum_{j=1}^{n} \frac{E[\varphi(Y_j)] - E[\varphi(Y_{j-1})]}{\varphi(u_j)}.$$

(If $\varphi(\xi) = \xi^2$, $u_1 = \cdots = u_n = \lambda$, we obtain Kolmogorov's inequality.)

Hint: Let I_j be 1 if $\{|Y_j| \leq u_j\}$ and 0 otherwise. Then

$$\begin{aligned}
\Pr\{|Y_j| \leq u_j;\ 1 \leq j \leq n\} &= E\left[\prod_{j=1}^{n} I_j\right] \\
&\geq E\left[\prod_{j=1}^{n-1} I_j \left(1 - \frac{\varphi(Y_n)}{\varphi(u_n)}\right)\right] \\
&\geq E\left[\prod_{j=1}^{n-1} I_j \left(1 - \frac{\varphi(Y_{n-1})}{\varphi(u_{n-1})}\right)\right] - \frac{E[\varphi(Y_n)] - E[\varphi(Y_{n-1})]}{\varphi(u_n)},
\end{aligned}$$

using successively that $\{\varphi(Y_n)\}$ is a submartingale, and $\{u_n\}$ increasing. Repeat.

39. Let $\{Y_n\}$ be a nonnegative submartingale and suppose b_n is a nonincreasing sequence of positive numbers. Suppose $\sum_{n=1}^{\infty}(b_n - b_{n+1})E[Y_n] < \infty$. Prove that

$$\lambda \Pr\{\sup_{k \geq 1} b_k Y_k > \lambda\} < \sum_{k=1}^{\infty} (b_k - b_{k+1})E[Y_k].$$

40. Let $\{X_n\}$ be a family of r.v.'s and let $\varphi(\xi)$ be a positive function defined for $\xi > 0$ satisfying

$$\frac{\varphi(\xi)}{\xi} \to \infty \qquad \text{as} \quad \xi \to \infty.$$

Suppose that

$$\sup_{m \geq 1} E[\varphi(|X_m|)] \leq K < \infty.$$

Show that $\{X_n\}$ is uniformly integrable.

41. Let $R_k(x)$ be a Rademacher functions $R_k(x) = \text{sign} \sin(2^{k+1}\pi x)$. Define

$$L_n(x) = \prod_{k=1}^{n} (1 + a_k R_k(x)) \qquad \text{for} \quad 0 \leq x \leq 1,$$

where a_k are constants.

Let $\mathscr{F}_n$, $n = 0, 1, \ldots$ be the field of sets induced by the partition of the unit interval, viz.,

$$\left(\frac{k}{2^{n+1}}, \frac{k+1}{2^{n+1}}\right) \qquad k = 0, 1, \ldots, 2^{n+1}.$$

Let $\mu(\text{dx})$ be a probability measure which assigns probability $1/2^{n+1}$ to each basic subinterval of $\mathscr{F}_n$. Show that

(1) L_n is a r.v. measurable $\mathscr{F}_n$.

(2) L_n is a martingale adapted to the σ-fields $\mathscr{F}_n$.

NOTES

Doob [1] developed martingale theory and demonstrated the broad usefulness of the concept.

REFERENCES

1. J. L. Doob, "Stochastic Processes." Wiley, New York, 1953.
2. Paul-André Meyer, "Martingales and Stochastic Integrals." Springer-Verlag, Berlin, New York, 1972 (Lecture Notes in Mathematics No. 284).
3. J. Neveu, "Mathematical Foundations of the Calculus of Probability." Holden-Day, San Francisco, 1965.
4. J. Neveu, "Martingales à Temps Discret." Masson, Paris, 1972.

Chapter 7

BROWNIAN MOTION

R. Brown, in 1827, observed that small particles immersed in a liquid exhibit ceaseless irregular motions. Historically, the Brownian motion process that is the subject of this chapter arose as an early attempt to explain this phenomenon. Today, the Brownian motion process and its many generalizations and extensions arise in numerous and diverse areas of pure and applied science such as economics, communication theory, biology, management science, and mathematical statistics.

The first four sections of this chapter provide an introduction that should be included in every first course in stochastic processes. The next section uses martingale methods to compute a number of expectations and probabilities associated with Brownian motion. It requires Section 5 of Chapter 6 as a prerequisite. The last sections treat more specialized topics.

1: Background Material

Certain special classes of stochastic processes have undergone extensive mathematical development. The Brownian motion process is the most renowned and historically the first which was thoroughly investigated. We will present a bare introduction to some of its salient features and hope thereby to whet the appetite of the reader for the elegant and elaborate theory of this process.

As a physical phenomenon the Brownian motion was discovered by the English botanist Brown in 1827. A mathematical description of this phenomenon was first derived from the laws of physics by Einstein in 1905. Since then the subject has made considerable progress. The physical theory was further perfected by Smoluchowski, Fokker, Planck, Burger, Furth, Ornstein, Uhlenbeck, Chandrasekhar, Kramers, and others. The mathematical theory was slower in developing because the exact mathematical description of the model posed difficulties, whereas some of the questions to which the physicists sought answers on the basis of this model were quite simple and intuitive. Many of the answers were obtained in a heuristic way by Bachelier in his 1900 dissertation[1] whereas the first

[1] Louis Bachelier, "Théorie de la spéculation" (doctoral dissertation in mathematics, University of Paris, March 29, 1900), *Annales de l'Ecole Normale Supérieure*, *Ser.* 3, **17**, 21–86 (1900). English translation: pp. 17–75 of P. H. Cootner (ed.) The Random Character of Stock Market Prices, MIT Press, Cambridge, Massachusetts, 1964.

concise mathematical formulation of the theory was given by Wiener in his 1918 dissertation and later papers. (See the References at the close of the chapter.)

In terms of our general framework of stochastic processes, the Brownian motion process is an example of a continuous time, continuous state space, Markov process.

Let $X(t)$ be the x component (as a function of time) of a particle in Brownian motion (cf. p. 21). Let x_0 be the position of the particle at time t_0, i.e., $X(t_0) = x_0$. Let $p(x, t|x_0)$ represent the conditional probability density of $X(t + t_0)$, given that $X(t_0) = x_0$. We postulate that the probability law governing the transitions is stationary in time and therefore $p(x, t|x_0)$ does not depend on the initial time t_0.

Since $p(x, t|x_0)$ is a density function in x, we have the properties

$$p(x, t|x_0) \geq 0, \qquad \int_{-\infty}^{\infty} p(x, t|x_0)\, dx = 1. \tag{1.1}$$

Further, we stipulate that, for small t, $X(t + t_0)$ is likely to be near $X(t_0) = x_0$. This is done formally by requiring

$$\lim_{t \to 0} p(x, t|x_0) = 0, \qquad \text{for } x \neq x_0. \tag{1.2}$$

From physical principles Einstein showed that $p(x, t|x_0)$ must satisfy the partial differential equation

$$\frac{\partial p}{\partial t} = D \frac{\partial^2 p}{\partial x^2}. \tag{1.3}$$

This is called the diffusion equation, and D is the diffusion coefficient. Small particles execute Brownian motion owing to collisions with the molecules in the gas or liquid in which they are suspended. The evaluation of D is based on the formula $D = 2RT/Nf$, where R is the gas constant, T is the temperature, N is Avogadro's number, and f is a coefficient of friction. By choosing the proper scale we may take $D = \frac{1}{2}$. Then we can verify directly that

$$p(x, t|x_0) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2t}(x - x_0)^2\right) \tag{1.4}$$

is a solution of (1.3), in fact, the unique solution under the boundary conditions (1.1) and (1.2). (The problem of the uniqueness of solutions of (1.3) needs to be formulated precisely and its analysis entails considerable care beyond the scope of this book.

Another approach to (1.3) is an approximation by means of a discrete random walk. Consider the symmetric random walk on the integers (see Example B, Section 2 of Chapter 2). Let $p_k(n)$ be the probability that a particle in this random walk finds itself k steps to the right of its starting point at time n. The Chapman–Kolmogorov relation [formula (3.2) of Chapter 2] for this process becomes

$$p_k(n+1) = \tfrac{1}{2}p_{k+1}(n) + \tfrac{1}{2}p_{k-1}(n),$$

which we may write as

$$p_k(n+1) - p_k(n) = \tfrac{1}{2}[p_{k+1}(n) - 2p_k(n) + p_{k-1}(n)]. \tag{1.5}$$

We recognize on the left the discrete version of the time derivative and on the right one half of the discrete version of the second derivative in the spatial variable. By an appropriate limiting process where the time between transitions shrinks to zero and simultaneously the size of the steps contracts appropriately to zero we may pass from (1.5) to (1.3).

Specifically, let the length of time between transitions be Δ, and the length of each step η. Then the analog of (1.5) is

$$\frac{p_{k\eta}((n+1)\Delta) - p_{k\eta}(n\Delta)}{\Delta} = \frac{\tfrac{1}{2}[p_{(k+1)\eta}(n\Delta) - 2p_{k\eta}(n\Delta) + p_{(k-1)\eta}(n\Delta)]}{\Delta}. \tag{1.6}$$

Now let Δ and η shrink to zero, preserving the relationship $\Delta = \eta^2$, and at the same time let n and k increase to ∞ so that $k\eta \to x$ while $n\Delta \to t$. Then $p_{k\eta}(n\Delta) \to p(x, t|0)$ and (1.6) passes formally into (1.3).

We will not attempt to rigorize this procedure. It is simple in concept, but requires rather delicate analysis to make it precise.

Another kind of limiting process for $p_k(n)$ requires the central limit theorem. We write

$$p_k(n) = \Pr\{X_1 + X_2 + \cdots + X_n = k\},$$

where $\{X_i\}$ represent the successive outcomes of tossing a fair coin, (i.e., $X_i = 1$ if heads and $X_i = -1$ if tails, each occurring with probability $\frac{1}{2}$). By the central limit theorem (see Section 1, Chapter 1).

$$\lim_{n\to\infty} \sum_{k=-\infty}^{\sqrt{n}x} p_k(n) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-u^2/2)\, du. \tag{1.7}$$

The limiting relation of (1.6) and that of (1.7) are essentially the same and are connected by the "invariance principle of stochastic processes." These heuristics can be made precise but are beyond the scope of this book.

2: *Joint Probabilities for Brownian Motion*

The transition probability density function (1.4) gives merely the probability distribution of $X(t) - X(0)$. The complete description of the Brownian motion process is furnished by the following definition.

Definition 2.1. *Brownian motion is a stochastic process* $\{X(t);\ t \geq 0\}$ *with the following properties:*

(a) *Every increment* $X(t+s) - X(s)$ *is normally distributed with mean* 0 *and variance* $\sigma^2 t$; σ *is a fixed parameter.*

(b) *For every pair of disjoint time intervals* $[t_1, t_2]$, $[t_3, t_4]$, *say* $t_1 < t_2 \leq t_3 < t_4$, *the increments* $X(t_4) - X(t_3)$ *and* $X(t_2) - X(t_1)$ *are independent random variables with distributions given in* (a), *and similarly for n disjoint time intervals where n is an arbitrary positive integer.*

(c) $X(0) = 0$ *and* $X(t)$ *is continuous at* $t = 0$.

This means that we postulate that a displacement $X(t+s) - X(s)$ is independent of the past, or alternatively, if we know $X(s) = x_0$, then no further knowledge of the values of $X(\tau)$ for $\tau < s$ has any effect on our knowledge of the probability law governing $X(t+s) - X(s)$. Written formally, this says that if $t > t_0 > t_1 > \cdots > t_n$,

$$\begin{aligned} &\Pr[X(t) \leq x | X(t_0) = x_0, X(t_1) = x_1, \cdots, X(t_n) = x_n] \\ &\quad = \Pr[X(t) \leq x | X(t_0) = x_0]. \end{aligned} \tag{2.1}$$

This is a statement of the Markov character of the process. We emphasize, however, that the independent increment assumption (b) is actually more restrictive than the Markov property.

Under the condition that $X(0) = 0$, the variance of $X(t)$ is $\sigma^2 t$. Hence σ^2 is sometimes called the variance parameter of the process. The process $\tilde{X}(t) = X(t)/\sigma$ is a Brownian motion process having a variance parameter of one, called *standard Brownian motion*. By this device we may always reduce an arbitrary Brownian motion to a standard Brownian motion; for the most part we derive results only for the latter.

By part (a) of the definition with $\sigma^2 = 1$, we have

$$\begin{aligned} \Pr[X(t) \leq x | X(t_0) = x_0] &= \Pr[X(t) - X(t_0) \leq x - x_0] \\ &= \frac{1}{\sqrt{2\pi(t-t_0)}} \int_{-\infty}^{x-x_0} \exp\left[-\frac{\alpha^2}{2(t-t_0)}\right] d\alpha. \end{aligned} \tag{2.2}$$

The consistency of part (b) of the definition with part (a) follows from well-known properties of the normal distribution, e.g., if $t_1 \leq t_2 \leq t_3$ then

$$X(t_3) - X(t_1) = [X(t_3) - X(t_2)] + [X(t_2) - X(t_1)].$$

On the right we have independent normal random variables with means 0 and variances $t_3 - t_2$ and $t_2 - t_1$, respectively. Hence their sum is normal with mean 0 and variance $t_3 - t_1$ as it should be.

It is not difficult using (2.1) and (2.2) to derive the joint density of $X(t_1), X(t_2), \ldots, X(t_n)(t_1 < t_2 < \cdots < t_n)$ subject to the condition $X(0) = 0$. Indeed, we only have to know the probability density of $X_1 = X(t_1) = x_1$, of $X_2 - X_1 = x_2 - x_1$, etc., and finally that of $X_n - X_{n-1} = x_n - x_{n-1}$. By part (b) of the definition we get at once the following expression for the density function:

$$f(x_1, \ldots, x_n) = p(x_1, t_1)p(x_2 - x_1, t_2 - t_1) \cdot \cdots \cdot p(x_n - x_{n-1}, t_n - t_{n-1}), \tag{2.3}$$

where

$$p(x, t) = \frac{1}{\sqrt{2\pi t}} \exp(-x^2/2t). \tag{2.4}$$

With the explicit formula (2.3) in hand we can compute in principle any set of conditional probabilities desired.

According to the Markov property, we know that, if $t_1 < t_2 < t_3$, the conditional density of $X(t_3)$, given $X(t_1)$ and $X(t_2)$, is the same as that of $X(t_3)$ given just $X(t_2)$.

However, the density of $X(t_2)$ given $X(t_1)$ and $X(t_3)$ is also of interest. Suppose, for definiteness, $X(t_1) = X(t_3) = 0$ and say specifically that $t_1 = 0$, $t_3 = 1$, and $t_2 = t(0 < t < 1)$.

By (2.3) the joint density of $X(t)$ and $X(1)$ is

$$f(x, y) = \frac{1}{2\pi\sqrt{t(1-t)}} \exp\left[-\frac{1}{2}\left\{\frac{x^2}{t} + \frac{(y-x)^2}{1-t}\right\}\right].$$

It follows that the conditional density of $X(t)$ given $X(0) = X(1) = 0$, denoted by $f_t(x|X(0) = X(1) = 0)$ is

$$\frac{1}{\sqrt{2\pi t(1-t)}} \exp\left[-\frac{1}{2}\frac{x^2}{t(1-t)}\right] \qquad -\infty < x < \infty.$$

In particular $E_c(X(t)) = 0$ and $E_c(X^2(t)) = t(1-t)$, where E_c refers to expectations taken under the conditions $X(0) = X(1) = 0$. The same methods yield the more general interpolation result.

Theorem 2.1. *The conditional density of $X(t)$ for $t_1 < t < t_2$ given $X(t_1) = A$ and $X(t_2) = B$ is a normal density with mean*

$$A + \frac{B - A}{t_2 - t_1}(t - t_1) \qquad \textit{and variance} \quad \frac{(t_2 - t)(t - t_1)}{t_2 - t_1}.$$

This can be reduced to the preceding case as follows. The conditional random variable $X(t)$ as indicated, i.e., the r.v. $X(t)$ subject to the conditions $X(t_1) = A$ and $X(t_2) = B$, has the same density as the random variable $A + X(t - t_1)$ under the condition $X(0) = 0$, $X(t_2 - t_1) = B - A$. This clearly has the same density as the random variable

$$A + X(t - t_1) + \frac{(t - t_1)}{t_2 - t_1}(B - A)$$

under the conditions $X(0) = 0$ and $X(t_2 - t_1) = 0$.

3: *Continuity of Paths and the Maximum Variables*

The physical origins of the Brownian motion process suggest that the possible realizations $X(t)$, as the graphs of the x coordinate of the position of a particle (i.e., the sample paths) whose movements result from continuous collisions in the surrounding medium are continuous functions. This fact is correct but a rigorous proof requires rather delicate analysis and is deferred until Section 7.

The sample paths $X(t)$, although continuous, are very kinky and their derivative exists nowhere. This fact is also rather deep. A complete description of the path structure of the Brownian motion process can be found in P. Lévy and in Ito and McKean (see References at the end of this chapter).

Using the property of the continuity of paths we will show how to calculate various interesting probability expressions of the Brownian motion. The first computation illustrates the use of the so-called *reflection principle.*

Bearing in mind the continuity of $X(t)$, we consider the collection of sample paths $X(t)$, $0 \leq t \leq T$, $X(0) = 0$, with the property that $X(T) > a$ $(a > 0)$. Since $X(t)$ is continuous and $X(0) = 0$ there exists a time τ (itself a random variable depending on the particular sample path) at which $X(t)$ first attains the value a.

For $t > \tau$, we reflect $X(t)$ about the line $x = a$ to obtain

$$\tilde{X}(t) = \begin{cases} X(t) & \text{for} \quad t < \tau, \\ a - [X(t) - a] & \text{for} \quad t > \tau \end{cases}$$

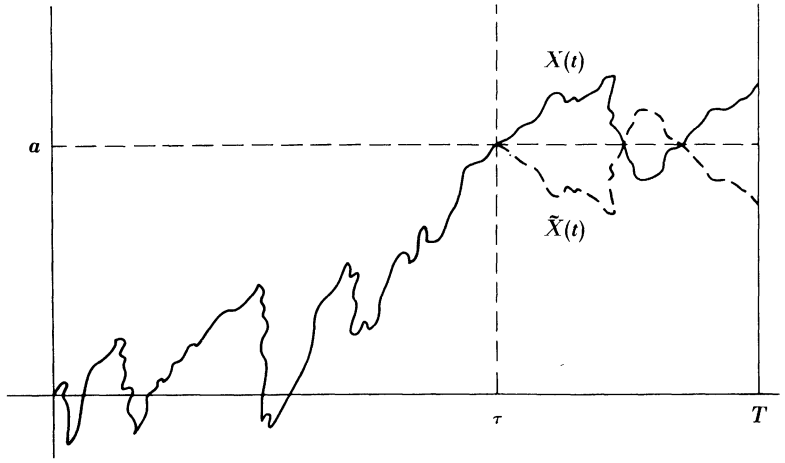


FIG. 1

(see Fig. 1). Note that $\tilde{X}(T) < a$ since $X(T) > a$. Because the probability law of the path for $t > \tau$, given $X(\tau) = a$, is symmetrical with respect to the values $x > a$ and $x < a$ and independent of the history prior to time τ, the reflection argument displays for every sample path with $X(T) > a$ two sample paths $X(t)$ and $\tilde{X}(t)$ with the same probability of occurrence such that

$$\max_{0 \le u \le T} X(u) \ge a \qquad \text{and} \qquad \max_{0 \le u \le T} \tilde{X}(u) \ge a.$$

Conversely, by the nature of this correspondence every sample function $X(t)$ for which $\max_{0 \le u \le T} X(u) \ge a$ results from either of two sample functions $X(t)$ with equal probability, one of which is such that $X(T) > a$, unless $X(T) = a$. But $\Pr\{X(T) = a\} = 0$. Thus, we may conclude, under the condition $X(0) = 0$, that

$$\Pr\left\{\max_{0 \le u \le T} X(u) \ge a\right\} = 2 \Pr\{X(T) > a\} = \frac{2}{\sqrt{2\pi T}} \int_a^\infty \exp(-x^2/2T)\, dx. \tag{3.1}$$

The above argument cannot be considered complete although the method is typical of a great deal of the analysis underlying the study of Markov processes with continuous paths. (Such processes are called diffusion processes.) A rigorous treatment would involve using the strong Markov property on the Markov time (see Section 4, Chapter 14) corresponding to the event of first passage from the value 0 to the value a.

With the help of (3.1) we may determine the distribution of the first time of reaching $a > 0$ subject to the condition $X(0) = 0$. Let T_a denote the time at which $X(t)$ first attains the value a where $X(0) = 0$. Then clearly

$$\Pr\{T_a \leq t\} = \Pr\left\{\max_{0 \leq u \leq t} X(u) \geq a | X(0) = 0\right\}. \tag{3.2}$$

But according to (3.1)

$$\Pr\left\{\max_{0 \leq u \leq t} X(u) \geq a | X(0) = 0\right\} = 2 \Pr\{X(t) > a\} = \frac{2}{\sqrt{2\pi t}} \int_a^\infty \exp\left[-\frac{1}{2}\frac{x^2}{t}\right] dx$$

and so

$$\Pr\{T_a \leq t\} = \sqrt{\frac{2}{\pi t}} \int_a^\infty \exp\left[-\frac{1}{2}\frac{x^2}{t}\right] dx.$$

The change of variable $x = y\sqrt{t}$ leads to

$$\Pr\{T_a \leq t\} = \sqrt{\frac{2}{\pi}} \int_{a/\sqrt{t}}^\infty \exp\left[-\frac{y^2}{2}\right] dy. \tag{3.3}$$

The density function of the random variable T_a is obtained by differentiating (3.3) with respect to t. Thus

$$f_{T_a}(t | X(0) = 0)\, dt = \frac{a}{\sqrt{2\pi}} t^{-3/2} \exp\left[-\frac{a^2}{2t}\right] dt. \tag{3.4}$$

Because of the symmetry and spatial homogeneity of the Brownian motion process we infer for the distribution (3.4) that

$$\begin{aligned}
&\Pr\left\{\min_{0 \leq u \leq t} X(u) \leq 0 | X(0) = a\right\} \\
&= \Pr\left\{\max_{0 \leq u \leq t} X(u) \geq 0 | X(0) = -a\right\} && \text{(by symmetry)} \\
&= \Pr\left\{\max_{0 \leq u \leq t} X(u) \geq a | X(0) = 0\right\} = \Pr\{T_a \leq t\} && \text{(by homogeneity)} \\
&= \frac{a}{\sqrt{2\pi}} \int_0^t u^{-3/2} \exp\left[-\frac{a^2}{2u}\right] du, && a > 0. \qquad (3.5)
\end{aligned}$$

Another way to express the result of (3.5) is as follows: If $X(t_0) = a$ then the probability $P(a)$ that $X(t)$ has at least one zero between t_0 and t_1 is

$$P(a) = \frac{|a|}{\sqrt{2\pi}} \int_0^{t_1 - t_0} \exp\left[-\frac{a^2}{2u}\right] u^{-3/2}\, du. \tag{3.6}$$

With this in hand we can calculate the probability α that if $X(0) = 0$ then $X(t)$ vanishes at least once in the interval (t_0, t_1).

In fact, we condition on the possible values of $X(t_0)$. Thus, if $X(t_0) = a$ then the probability that $X(t)$ vanishes in the interval (t_0, t_1) is $P(a)$. By the law of total probabilities

$$\alpha = \int_0^\infty P(a) \Pr\{|X(t_0)| = a \Big| X(0) = 0\}\, da = \sqrt{\frac{2}{\pi t_0}} \int_0^\infty P(a) \exp\left[-\frac{a^2}{2t_0}\right] da. \tag{3.7}$$

Substituting from (3.6) and then interchanging the order of integration yields

$$\begin{aligned} \alpha &= \sqrt{\frac{2}{\pi t_0}} \int_0^\infty \exp\left[-\frac{a^2}{2t_0}\right] \frac{a}{\sqrt{2\pi}} \left(\int_0^{t_1 - t_0} \exp\left[-\frac{a^2}{2u}\right] u^{-3/2}\, du \right) da \\ &= \frac{1}{\pi\sqrt{t_0}} \int_0^{t_1 - t_0} u^{-3/2} \left(\int_0^\infty a \exp\left[-\frac{a^2}{2}\left(\frac{1}{u} + \frac{1}{t_0}\right)\right] da \right) du. \end{aligned} \tag{3.8}$$

The inner integral can be integrated exactly and after simplifying we get

$$\alpha = \frac{\sqrt{t_0}}{\pi} \int_0^{t_1 - t_0} \frac{du}{(t_0 + u)\sqrt{u}}.$$

The change of variables $u = t_0 v^2$ produces

$$\alpha = \frac{2}{\pi} \int_0^{\sqrt{(t_1 - t_0)/t_0}} \frac{dv}{1 + v^2} = \frac{2}{\pi} \arctan \sqrt{\frac{t_1 - t_0}{t_0}},$$

which we may write by virtue of some standard trigonometric relations in the form

$$\sqrt{\frac{t_0}{t_1}} = \cos\frac{\pi\alpha}{2} \qquad \text{or} \qquad \alpha = \frac{2}{\pi} \arccos \sqrt{\frac{t_0}{t_1}}.$$

To sum up, we have

Theorem 3.1. *The probability that $X(t)$ has at least one zero in the interval (t_0, t_1), given $X(0) = 0$, is*

$$\alpha = \frac{2}{\pi} \arccos \sqrt{\frac{t_0}{t_1}}.$$

With the aid of the same "reflection principle" we now solve the following problem: Determine

$$A_t(x, y) = \Pr\left\{ X(t) > y, \min_{0 \le u \le t} X(u) > 0 \Big| X(0) = x \right\} \tag{3.9}$$

for $x > 0$ and $y > 0$. To determine (3.9), we start with the obvious relation

$$\Pr\{X(t) > y | X(0) = x\}$$
$$= A_t(x, y) + \Pr\left\{X(t) > y, \min_{0 \le u \le t} X(u) \le 0 | X(0) = x\right\}. \tag{3.10}$$

The reflection principle is applied to the last term.

Figure 2 is the appropriate picture to guide the analysis; we may deduce that

$$\Pr\{X(t) > y, \min_{0 \le u \le t} X(u) \le 0 | X(0) = x\}$$
$$= \Pr\left\{X(t) < -y, \min_{0 \le u \le t} X(u) \le 0 | X(0) = x\right\} \tag{3.11}$$
$$= \Pr\{X(t) < -y | X(0) = x\}.$$

The reasoning behind (3.11) goes as follows: Consider a path starting at $x > 0$ satisfying $X(t) > y$ which reaches 0 at some intermediate time τ.

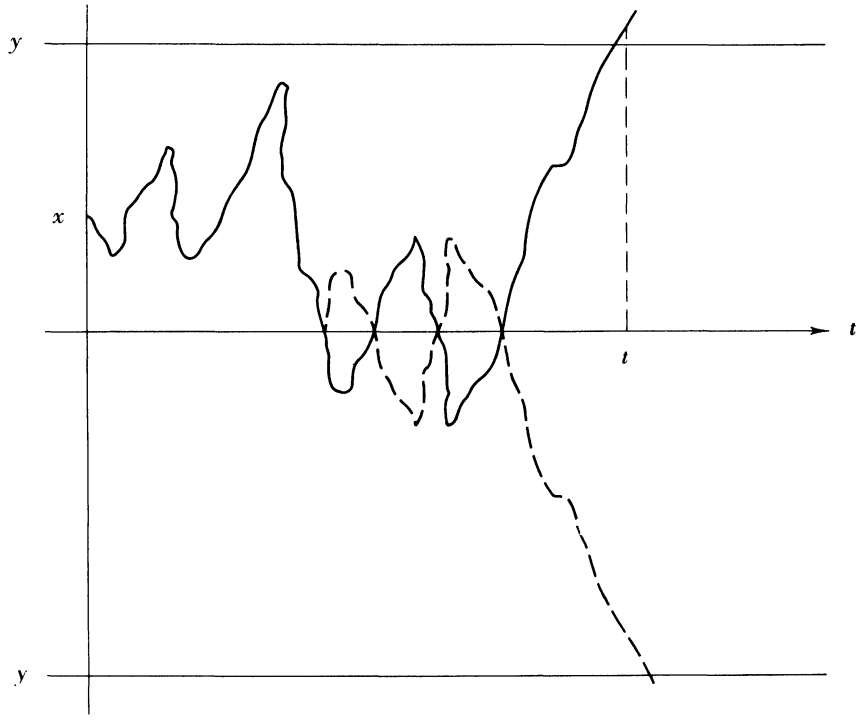


FIG. 2

By reflecting such a path about zero after time τ we obtain a path starting from x and reaching a value smaller than $-y$ at time t. This implies the equality of the first two terms of (3.11). The equality of the last two terms is clear from their meaning since the condition $\min_{0\le u\le t} X(u) \le 0$ is superfluous in view of the requirement $X(t) < -y$ $(y > 0)$. Inserting (3.11) in (3.10) yields

$$\begin{aligned} A_t(x, y) &= \Pr\{X(t) > y | X(0) = x\} - \Pr\{X(t) < -y | X(0) = x\} \\ &= \Pr\{X(t) > y - x | X(0) = 0\} - \Pr\{X(t) < -(y + x) | X(0) = 0\} \\ &\qquad \text{(by spatial homogeneity)} \\ &= \Pr\{X(t) > y - x | X(0) = 0\} - \Pr\{X(t) > y + x | X(0) = 0\} \\ &\qquad \text{(by symmetry)} \\ &= \int_{y-x}^{y+x} p(u, t)\, du, \end{aligned} \tag{3.12}$$

where $p(u, t) = (2\pi t)^{-1/2} \exp\{-u^2/2t\}$ is the transition probability density function for the Brownian motion process.

As a final application of the reflection principle we now derive the joint probability density function for

$$M(t) = \max_{0\le u\le t} X(u), \qquad \text{and} \qquad Y(t) = M(t) - X(t).$$

The reflection principle, with Fig. 3 as an aid, implies

$$\begin{aligned} \Pr\{M(t) \ge m, X(t) \le x\} &= \Pr\{X(t) \ge 2m - x\} \\ &= 1 - \Phi\left(\frac{2m - x}{\sqrt{t}}\right), \qquad m \ge 0, \quad m \ge x, \end{aligned}$$

where $\Phi(x)$ is the distribution function of the standard normal density $\phi(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$. Differentiate with respect to x and then with respect to m, changing the sign, to get the joint density function for $M(t)$ and $X(t)$.

The calculations are:

$$\begin{aligned} -\frac{d}{dm}\frac{d}{dx}\left\{1 - \Phi\left(\frac{2m - x}{\sqrt{t}}\right)\right\} &= -\frac{d}{dm}\left\{\frac{1}{\sqrt{t}}\phi\left(\frac{2m - x}{\sqrt{t}}\right)\right\} \\ &= \frac{2m - x}{t}\frac{2}{\sqrt{t}}\phi\left(\frac{2m - x}{\sqrt{t}}\right), \end{aligned}$$

using the elementary relation

$$\frac{d}{dx}\phi(x) = -x\phi(x).$$

Denoting this joint density by $f(m, x)$, we have explicitly

$$f(m, x) = \sqrt{\frac{2}{\pi t^3}}\,(2m - x)\exp[-(2m - x)^2/2t], \qquad \begin{cases} 0 \le m, \\ x \le m. \end{cases} \tag{3.13}$$

To obtain the joint density $g(m, y)$ of $M(t)$, $Y(t) = M(t) - X(t)$, we have

$$\Pr\{M(t) \le a,\ Y(t) \le b\} = \int_{m \le a} \int_{y \le b} f(m, m - y)\,dy\,dm,$$

so the desired joint density is $g(m, y) = f(m, m - y)$ or

$$g(m, y) = \sqrt{\frac{2}{\pi t^3}}\,(m + y)\exp[-(m + y)^2/2t], \qquad m \ge 0, \quad y \ge 0. \tag{3.14}$$

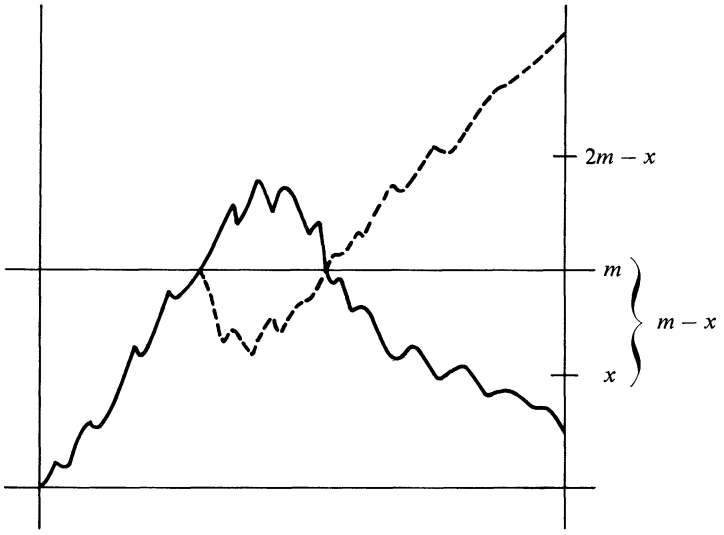


FIG. 3

4: Variations and Extensions

We claim that if $X(t)$ is a standard Brownian motion process, then the processes

$$X_1(t) = cX(t/c^2), \qquad \text{for fixed} \quad c > 0,$$

$$X_2(t) = \begin{cases} tX(1/t), & \text{for} \quad t > 0, \\ 0, & \text{for} \quad t = 0, \end{cases}$$

and

$$X_3(t) = X(t + h) - X(h), \qquad \text{for fixed} \quad h > 0.$$

are each a version of standard Brownian motion.

We check the requirements of Definition 2.1. Manifestly, every increment $X_i(t+s) - X_i(t)$ in these processes is normally distributed with zero mean, and the increments over disjoint time intervals manifestly determine independent r.v's. To continue the verification that these are Brownian motions we need the variance $E[\{X_i(s+t) - X_i(s)\}^2]$ in each case. We have

$$\begin{aligned} E[\{X_1(t+s) - X_1(s)\}^2] &= c^2 E[\{X((t+s)/c^2) - X(s/c^2)\}^2] \\ &= c^2\{(t+s)/c^2 - s/c^2\} = t, \end{aligned}$$

$$\begin{aligned} E[\{X_2(t+s) - X_2(s)\}^2] &= E\left[\left\{sX\left(\frac{1}{s}\right) - (t+s)X\left(\frac{1}{t+s}\right)\right\}^2\right] \\ &= s^2 E\left[\left\{X\left(\frac{1}{s}\right) - X\left(\frac{1}{t+s}\right)\right\}^2\right] + t^2 E\left[\left\{X\left(\frac{1}{t+s}\right)\right\}^2\right] \\ &= s^2\left\{\frac{1}{s} - \frac{1}{t+s}\right\} + t^2\,\frac{1}{t+s} \\ &= t, \end{aligned}$$

and

$$E[\{X_3(t+s) - X_3(s)\}^2] = E[\{X(t+h+s) - X(h+s)\}^2] = t.$$

To complete the analysis, it is necessary to check that each $X_i(t)$ is continuous at the origin. This property is obvious for $X_1(t)$ and $X_3(t)$ but needs some arguments for $X_2(t)$. Equivalently, in the latter case it is enough to show that

$$\Pr\left\{\lim_{t\to\infty} \frac{X(t)}{t} = 0 \,|\, X(0) = 0\right\}.$$

(See Problems 14 and 15.)

Other modifications of Brownian motion lead to new and important stochastic processes. Here are some examples.

A. BROWNIAN MOTION REFLECTED AT THE ORIGIN

Let $\{X(t),\ t \geq 0\}$ be a Brownian motion process. A stochastic process having the distribution of

$$\begin{aligned} Y(t) &= |X(t)|, \qquad t \geq 0 \\ &= \begin{cases} X(t), & \text{if } X(t) \geq 0, \\ -X(t), & \text{if } X(t) < 0, \end{cases} \end{aligned}$$

is called *Brownian motion reflected at the origin*, which we abbreviate to *reflected Brownian motion*.

Because of the spatial symmetry of Brownian motion, reflected Brownian motion is Markov. Indeed, in the case of standard Brownian motion ($\sigma^2 = 1$),

$$\begin{aligned}
&\Pr\{Y(t_n + s) \le z \,|\, Y(t_0) = x_0, \ldots, Y(t_n) = x_n\} \\
&\quad = \Pr\{-z \le X(t_n + s) \le +z \,|\, X(t_0) = \pm x_0, \ldots, X(t_n) = \pm x_n\} \\
&\quad = \Pr\{-z \le X(t_n + s) \le +z \,|\, X(t_0) = x_0, \ldots, X(t_n) = x_n\} \\
&\qquad\qquad \text{by symmetry} \\
&\quad = \Pr\{-z \le X(t_n + s) \le +z \,|\, X(t_n) = x_n\} \\
&\quad = \int_{-z}^{+z} p(y - x_n, s)\, dy, \qquad (4.1)
\end{aligned}$$

where

$$p(x, t) = \frac{1}{\sqrt{2\pi t}} \exp\{-x^2/2t\}, \qquad (4.2)$$

and $0 \le t_0 < t_1 < \cdots < t_n$, $s > 0$. Thus reflected Brownian motion furnishes an example of another continuous time Markov process whose sample paths $Y(t)$ are continuous. The state space consists of the set of nonnegative real numbers.

Since the moments of $Y(t)$ are the same as the moments of $|X(t)|$, the mean and variance of reflected Brownian motion may be computed easily. Under the condition $Y(0) = 0$, for example, we have,

$$\begin{aligned}
E[Y(t)] &= \int_{-\infty}^{+\infty} |x| p(x, t)\, dx \\
&= 2 \int_0^{\infty} \frac{x}{\sqrt{2\pi t}} \exp\{-x^2/2t\}\, dx \\
&= \sqrt{\frac{2t}{\pi}}.
\end{aligned}$$

The integral was evaluated through the change of variable $y = x/\sqrt{t}$. Also,

$$\begin{aligned}
\text{variance of } Y(t) &= E[Y(t)^2] - E[Y(t)]^2 \\
&= E[|X(t)|^2] - 2t/\pi \\
&= (1 - 2/\pi)t.
\end{aligned}$$

By changing variables in (4.1) it is easily seen that, for $t > 0$, $Y(t)$ is a continuous random variable with transition probability density function

$$p_t(x, y) = p(y - x, t) + p(y + x, t), \tag{4.3}$$

for which

$$\Pr\{\alpha \leq Y(t) \leq \beta | Y(0) = x\} = \int_\alpha^\beta p_t(x, y)\, dy,$$

where $p(x, t)$ is given in (4.2).

B. BROWNIAN MOTION ABSORBED AT THE ORIGIN

Suppose the initial value $X(0) = x$ of a Brownian motion process is positive† and let τ be the first time the process reaches zero. A stochastic process having the distribution of

$$Z(t) = \begin{cases} X(t), & \text{for } t \leq \tau, \\ 0 & \text{for } t > \tau, \end{cases} \tag{4.4}$$

is called *Brownian motion absorbed at the origin*, which we will shorten to *absorbed Brownian motion.*

Again, absorbed Brownian motion is a continuous time Markov process. We verify the Markov property in the form,

$$\begin{aligned} &\Pr\{Z(t_n + s) > y | Z(t_0) = x_0, \ldots, Z(t_{n-1}) = x_{n-1}, Z(t_n) = x\} \\ &\quad = \Pr\{Z(t_n + s) > z | Z(t_n) = x\}, \end{aligned}$$

where $x > 0$ and $0 < t_0 < \cdots < t_n$, $s > 0$. (The easier case, when $x = 0$ is left to the reader.) We compute this by way of a Brownian motion process $X(t)$ related to $Z(t)$ as in (4.1). The condition $x > 0$ entails

$$\min_{0 \leq u \leq t_n} X(u) > 0.$$

Hence

$$\begin{aligned} &\Pr\{Z(t_n + t) > y | Z(t_0) = x_0, \ldots, Z(t_{n-1}) = x_{n-1}, Z(t_n) = x\} \\ &\quad = \Pr\{Z(t_n + t) > y | Z(t_0) = x_0, \ldots, Z(t_{n-1}) = x_{n-1}, Z(t_n) \\ &\qquad = x, \min_{0 \leq v \leq t_n} X(u) > 0\} \\ &\quad = \Pr\{X(t_n + t) > y, \min_{0 \leq u \leq t} X(t_n + u) > 0 | X(t_0) = x_0, \ldots, X(t_{n-1}) \\ &\qquad = x_{n-1}, X(t_n) = x\} \\ &\quad = \Pr\{X(t) > y, \min_{0 \leq u \leq t} X(u) > 0 | X(0) = x\} \\ &\quad = A_t(x, y), \end{aligned}$$

† To define Brownian motion conditioned on $X(0) = x$, or "Brownian motion starting from x," replace "$X(0) = 0$" by "$X(0) = x$" in Part (c) of Definition 2.1.

where $A_t(x, y)$ was defined in (3.9) and computed in (3.12) to be

$$A_t(x, y) = \int_y^\infty [p(u - x, t) - p(u + x, t)]\, du$$
$$= \int_y^{y+2x} p(u - x, t)\, du.$$

Under the condition $Z(0) = x > 0$, $Z(t)$ is a random variable whose distribution has both discrete and continuous parts. The discrete part is

$$\Pr\{Z(t) = 0 | Z(0) = x\} = 1 - A_t(x, 0)$$
$$= 1 - \int_0^{2x} p(u - x, t)\, du.$$
$$= 1 - \int_{-x}^{x} p(u, t)\, du$$
$$= 2 \int_x^\infty p(u, t)\, du$$
$$= 2 \int_0^\infty p(u + x, t)\, du.$$

In the region $z > 0$, $Z(t)$ is a continuous random variable and for $0 < a < b$

$$\Pr\{a < Z(t) < b | Z(0) = x\} = A_t(x, a) - A_t(x, b)$$
$$= \int_a^b [p(u - x, t) - p(u + x, t)]\, du.$$

Thus the transition probability density function for the continuous part of absorbed Brownian motion is

$$p_t(x, y) = p(y - x, t) - p(y + x, t).$$

C. BROWNIAN MOTION WITH DRIFT

Let $\{\tilde{X}(t), t \geq 0\}$ be a Brownian motion process. *Brownian motion with drift* is a stochastic process having the distribution of

$$X(t) = \tilde{X}(t) + \mu t, \qquad t \geq 0,$$

where μ is a constant, called the *drift parameter*. Alternatively, we may describe Brownian motion with drift in a manner that parallels Definition 2.1.

Definition 4.1. *Brownian motion with drift is a stochastic process* $\{X(t);\, t \geq 0\}$ *with the following properties:*

(a) *Every increment* $X(t+s) - X(s)$ *is normally distributed with mean* μt *and variance* $\sigma^2 t$; μ, σ *being fixed constants.*

(b) *For every pair of disjoint time intervals* $[t_1, t_2]$, $[t_3, t_4]$, *say,* $t_1 < t_2 \leq t_3 < t_4$, *the increments* $X(t_4) - X(t_3)$ *and* $X(t_2) - X(t_1)$ *are independent random variables with distributions given in* (a), *and similarly for n disjoint time intervals, where n is an arbitrary positive integer.*

(c) $X(0) = 0$ *and* $X(t)$ *is continuous at* $t = 0$.

As before, it follows that a displacement $X(t+s) - X(s)$ is independent of the past, or alternatively, if we know $X(s) = x_0$, then no further knowledge of the values of $X(\tau)$ for $\tau < s$ affects the conditional probability law governing $X(t+s) - X(s)$. Written formally, this says that if $t > t_0 > t_1 > \cdots > t_n$,

$$\begin{aligned} &\Pr\{X(t) \leq x | X(t_0) = x_0,\, X(t_1) = x_1, \ldots, X(t_n) = x_n\} \\ &\quad = \Pr\{X(t) \leq x | X(t_0) = x_0\}. \end{aligned} \tag{4.5}$$

This is a statement of the Markov character of the process. We emphasize, however, that the independent increment assumption (b) is actually more restrictive than the Markov property. By part (a) of the definition, we have

$$\begin{aligned} \Pr\{X(t) \leq x | X(t_0) = x_0\} &= \Pr\{X(t) - X(t_0) \leq x - x_0\} \\ &= \int_{-\infty}^{x - x_0} \frac{1}{\sqrt{2\pi(t-t_0)}\sigma} \exp\left\{-\frac{(y - \mu(t-t_0))^2}{2(t-t_0)\sigma^2}\right\} dy \\ &= \int_{-\infty}^{\{x - x_0 - \mu(t-t_0)\}/\sigma} p(t - t_0, y)\, dy, \end{aligned}$$

where

$$p(t, x) = \frac{1}{\sqrt{2\pi t}} \exp\{-x^2/2t\}.$$

When $\mu \neq 0$, the process is no longer symmetric, and the reflection argument may not be used to compute the distribution of the maximum of the process. We will compute this distribution in the next section using facts of martingale theory.

D. GEOMETRIC BROWNIAN MOTION

Let $\{X(t),\ t \geq 0\}$ be a Brownian motion process with drift μ and diffusion coefficient σ^2. The process defined by

$$Y(t) = e^{X(t)}, \qquad t \geq 0,$$

is sometimes called *geometric Brownian motion*. The state space is the interval $(0, \infty)$.

Since $Y(t) = Y(0)e^{X(t)-X(0)}$, using the characteristic function for the normal distribution, we compute

$$\begin{aligned} E[Y(t)|\, Y(0) = y] &= yE[e^{X(t)-X(0)}] \\ &= y \exp\{t(\mu + \tfrac{1}{2}\sigma^2)\}, \end{aligned}$$

and

$$\begin{aligned} E[Y(t)^2|\, Y(0) = y] &= y^2 E[e^{2[X(t)-X(0)]}] \\ &= y^2 \exp\{t[2\mu + \tfrac{1}{2}4\sigma^2]\}, \end{aligned}$$

so that the variance of $Y(t)$ is

$$\begin{aligned} \operatorname{Var}[Y(t)|\, Y(0) = y] &= E[Y(t)^2|\, Y(0) = y] - \{E[Y(t)|\, Y(0) = y]\}^2 \\ &= y^2\{\exp[2t(\mu + \tfrac{1}{2}2\sigma^2)] - \exp[2t(\mu + \tfrac{1}{2}\sigma^2)]\} \\ &= y^2 \exp[2t(\mu + \tfrac{1}{2}\sigma^2)][\exp(t\sigma^2) - 1]. \end{aligned}$$

5: Computing Some Functionals of Brownian Motion by Martingale Methods

A number of important quantities can be expeditiously calculated by applying the optional sampling theorem to martingales associated with Brownian motion. We remarked in Chapter 1 that a standard Brownian motion process $\{X(t);\ t \geq 0\}$ is a martingale, but this is by no means the only martingale of interest in this context. Both of the processes

$$U(t) = X^2(t) - t,$$

and (5.1)

$$V(t) = \exp\{\lambda X(t) - \tfrac{1}{2}\lambda^2 t\},$$

where λ is an arbitrary real constant, are martingales with respect to standard Brownian motion $\{X(t)\}$. We present direct validation. Indeed,

we have, for $0 \leq t_1 \leq \cdots \leq t_n = t$, and $s > 0$,

$$\begin{aligned} E[U(t+s)|X(t_1), \ldots, X(t_n)] &= E[X^2(t+s)|X(t)] - (t+s) \\ &= E[\{X(t+s) - X(t)\}^2|X(t)] \\ &\quad + 2E[X(t)\{X(t+s) - X(t)\}|X(t)] \\ &\quad + E[X^2(t)|X(t)] - (t+s) \\ &= s + 2 \times 0 + X^2(t) - (t+s) \\ &= U(t). \end{aligned}$$

Similarly,

$$\begin{aligned} E[V(t+s)|X(t_1), \ldots, X(t_n)] &= V(t) \times E[\exp\{\lambda[X(t+s) - X(t)] - \tfrac{1}{2}\lambda^2 s\}] \\ &= V(t). \end{aligned}$$

Digression. It is useful to place the martingale examples of (5.1) in a more facile framework. A general construction of martingales associated with a Markov process $\{X(t);\ t \geq 0\}$ having stationary transition probability density

$$p(t, x|y)\,dx = \Pr\{x \leq X(t) < x + dx|X(0) = y\}$$

runs as follows. Let $u(x, t)$ solve the functional equation

$$u(x, s) = \int p(t, \xi|x)u(\xi, t+s)\,d\xi, \qquad s, t > 0. \tag{5.2}$$

We claim that $Z(t) = u(X(t), t)$ determines a martingale adapted to the sigma fields $\mathcal{F}_t = \sigma\{X(u);\ 0 \leq u \leq t\}$ generated by the history of $X(t)$ up to time t, provided $E[|Z(t)|] < \infty$.

Proof. We compute

$$\begin{aligned} E[Z(t+s)|\mathcal{F}(s)] &= E[Z(t+s)|X(s)] \qquad \text{(by the Markov property)} \\ &= E[u(X(t+s), t+s)|X(s)] \\ &= \int p(t, \xi|X(s))u(\xi, t+s)d\xi \\ &= u(X(s), s) \qquad \text{(using the functional equation)} \\ &= Z(s). \end{aligned}$$

Thus, the martingale property for $\{Z(t)\}$ is fully corroborated.

A direct calculation reveals that

$$u(x, t) = x^2 - t \qquad \text{and} \qquad v(x, t) = \exp\{\lambda x - \tfrac{1}{2}\lambda^2 t\}$$

obey the functional relation (5.2) with

$$p(t, x|y) = \frac{1}{\sqrt{2\pi t}} \exp\{-(x-y)^2/2t\}.$$

In the case of the Brownian motion process it can be shown further that if $u(x, t)$ is sufficiently differentiable and fulfills the relation (5.2), then $u(x, t)$ also satisfies

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial^2 u}{\partial x^2} = 0. \tag{5.3}$$

In Chapter 15 on diffusion processes we provide a fuller proof of this last assertion. The converse is also correct, affirming that where (5.3) prevails, (5.2) also ensues.

Here are some sample calculations involving the martingales of (5.1). Let $a < 0 < b$ be given and let $T = T_{ab}$ be the first time the process reaches a or b:

$$T_{ab} = \inf\{t \geq 0 : X(t) = a \quad \text{or} \quad X(t) = b\},$$

and let $T \wedge n = \min\{T, n\}$.

Since $\{U(t)\}$ is a martingale, $E[U(T \wedge n)] = E[U(0)] = 0$, which gives

$$E[T \wedge n] = E[X^2(T \wedge n)] \leq (|a| + b)^2.$$

Thus

$$\begin{aligned} E[T] &= \lim_{n\to\infty} \int_0^n \Pr\{T > t\}\, dt \\ &= \lim_{n\to\infty} E[T \wedge n] \leq (|a| + b)^2. \end{aligned}$$

The important point is that $T = T_{ab}$ is finite. Even more, T_{ab} has a finite mean.

Let u be the probability that the $\{X(t)\}$ process reaches b before it reaches a, or

$$u = \Pr\{X(T_{ab}) = b\}.$$

Now $\{X(t)\}$ is a martingale, $T = T_{ab}$ is finite, and $\{X(t \wedge T)\}$ is bounded. Thus the optional stopping theorem applies and

$$\begin{aligned} 0 &= E[X(T)] \\ &= a \Pr\{X(T) = a\} + b \Pr\{X(T) = b\} \\ &= a[1 - u] + bu, \end{aligned}$$

so that

$$u = \Pr\{X(T_{ab}) = b\} = \frac{|a|}{|a| + b}.$$

We now return to the martingale $\{U(t)\}$, noticing $E[U(T)] = 0$, to compute

$$\begin{aligned} E[T_{ab}] &= E[X^2(T_{ab})] \\ &= a^2[1-u] + b^2 u \\ &= a^2 \frac{b}{|a|+b} + b^2 \frac{|a|}{|a|+b} \\ &= |a|b. \end{aligned}$$

By changing the origin and scale of a given process we may compute analogous quantities for a Brownian motion with variance parameter σ^2 and starting position $X(0) = x$. The result is:

Theorem 5.1. *Let $\{X(t); t \geq 0\}$ be a Brownian motion process with variance σ^2 and $X(0) = x$. Let a, b with $a < x < b$ be given, and let T be the first time the process reaches a or b. Then*

$$\Pr\{X(T) = b | X(0) = x\} = (x-a)/(b-a),$$

and

$$E[T|X(0) = x] = \frac{1}{\sigma^2}(b-x)(x-a).$$

Let us turn to a Brownian motion process $\{X(t); t \geq 0\}$ with drift $\mu \neq 0$ and variance σ^2. Then, for any real λ,

$$V(t) = \exp\{\lambda X(t) - (\lambda\mu + \tfrac{1}{2}\lambda^2\sigma^2)t\} \tag{5.4}$$

defines a martingale. Let us choose

$$\lambda_0 = -2\mu/\sigma^2,$$

so that the second term in the exponent of (5.4) vanishes. Then $V_0(t) = \exp\{\lambda_0 X(t)\}$ is a martingale. We apply the optional stopping theorem, with T_{ab} being the first time the process reaches $a < 0$ or $b > 0$, to learn

$$\begin{aligned} 1 &= E[V_0(T_{ab})] \\ &= \Pr\{X(T_{ab}) = a\} \exp\{\lambda_0 a\} + \Pr\{X(T_{ab}) = b\} \exp\{\lambda_0 b\}, \end{aligned}$$

and

$$\Pr\{X(T_{ab}) = b\} = \frac{1 - \exp\{\lambda_0 a\}}{\exp\{\lambda_0 b\} - \exp\{\lambda_0 a\}},$$

where $\lambda_0 = -2\mu/\sigma^2$. Again, we may translate the origin to treat the case $X(0) = x$.

Theorem 5.2. *Let $\{X(t);\ t \geq 0\}$ be a Brownian motion process with drift $\mu \neq 0$ and variance σ^2, and suppose $X(0) = x$. The probability that the process reaches the level $b > x$ before hitting $a < x$ is given by*

$$\Pr\{X(T_{ab}) = b | X(0) = x\} = \frac{\exp(-2\mu x/\sigma^2) - \exp(-2\mu a/\sigma^2)}{\exp(-2\mu b/\sigma^2) - \exp(-2\mu a/\sigma^2)}.$$

Corollary 5.1. *Let $X(t)$ be a Brownian motion process with drift $\mu < 0$. Let*

$$W = \max_{0 \leq t < \infty} X(t) - X(0).$$

Then W has the exponential distribution

$$\Pr\{W \geq w\} = e^{-\lambda w}, \qquad w \geq 0,$$

where $\lambda = 2|\mu|/\sigma^2$.

Proof. We let $a \to -\infty$ in the formula of Theorem 5.2. Since $\mu < 0$, $\exp(-\mu a/\sigma^2) \to 0$. Thus

$$\lim_{a \to -\infty} \Pr\{X(T_{ab}) = b | X(0) = x\} = \exp[2\mu(b - x)/\sigma^2].$$

But as $a \to -\infty$, the left-hand side becomes the probability that the process ever reaches b, that is, the probability that the maximum of the process ever exceeds b. Thus, for $w = b - x$,

$$\begin{aligned} \Pr\{W \geq w\} &= \Pr\left\{ \max_{0 \leq t < \infty} X(t) > b | X(0) = x \right\} \\ &= e^{-\lambda w}, \end{aligned}$$

with $\lambda = 2|\mu|/\sigma^2$, as claimed. ∎

As a last example, we calculate the Laplace transform of the first passage time to a single barrier. We let $z > 0$ be fixed and $T = T_z$ be the first time, if any, the process reaches the level z:

$$T = T_z = \begin{cases} \inf\{t : X(t) \geq z\}, & \text{if } X(t) \geq z, \text{ for some } t \geq 0, \\ \infty, & \text{if } X(t) < z, \text{ for all } t \geq 0. \end{cases}$$

Set $\theta = \lambda\mu + \frac{1}{2}\lambda^2\sigma^2$. Then $V(t) = \exp\{\lambda X(t) - \theta t\}$ is a martingale, and, if $X(0) = 0$,

$$1 = E[V(T \wedge t)],$$

or

$$1 = E[\exp\{\lambda X(T \wedge t) - \theta(T \wedge t)\}].$$

Let us suppose $\lambda \geq 0$ is sufficiently large to ensure $\theta \geq 0$. Then

$$0 \leq V(T \wedge t) \leq e^{\lambda z},$$

so that, using Lemma 3.3 of Chapter 6, we may pass to the limit as $t \to \infty$. We obtain

$$\lim_{t \to \infty} V(t \wedge T) = \begin{cases} \exp\{\lambda z - \theta T\}, & \text{if } T < \infty, \\ 0, & \text{if } T = \infty, \end{cases}$$

so that

$$\begin{aligned} 1 &= \lim_{t \to \infty} E[V(T \wedge t)] \\ &= e^{\lambda z} E[e^{-\theta T}], \end{aligned}$$

or

$$E[e^{-\theta T}] = e^{-\lambda z}.$$

It remains only to relate θ and λ. We have

$$\tfrac{1}{2}\sigma^2\lambda^2 + \mu\lambda - \theta = 0,$$

or

$$\lambda = \frac{-\mu \pm \sqrt{\mu^2 + 2\sigma^2\theta}}{\sigma^2}.$$

We require $\lambda \geq 0$, which implies

$$\lambda = \frac{1}{\sigma^2}(\sqrt{\mu^2 + 2\sigma^2\theta} - \mu).$$

When $\mu < 0$, T has a defective probability distribution, that is, T is infinite with positive probability, and

$$\begin{aligned} \Pr\{T < \infty\} &= \lim_{\theta \to 0} E[e^{-\theta T}] \\ &= \lim_{\theta \to 0} \exp\left[-\frac{z}{\sigma^2}(\sqrt{\mu^2 + 2\sigma^2\theta} - \mu)\right] \\ &= \exp(-2z|\mu|/\sigma^2), \end{aligned}$$

which agrees with Corollary 5.1. When $\mu \geq 0$, $T < \infty$ with certainty, and the Laplace transform is

$$E[e^{-\theta T}] = \exp\left[-\frac{z}{\sigma^2}(\sqrt{\mu^2 + 2\sigma^2\theta} - \mu)\right]. \tag{5.5}$$

For Brownian motion with drift $\mu \geq 0$, it is possible to invert the transform (5.5) and obtain an explicit expression for the probability density function of T_z. We satisfy ourselves with quoting the result.

Theorem 5.3. *Let $X(t)$ be a Brownian motion with drift $\mu \geq 0$. Let $z > X(0) = x$ be given and let T_z be the first time the process reaches the level z. Conditioned on $X(0) = x$, T_z has the probability density function*

$$f(t; x, z) = \frac{(z - x)}{\sigma\sqrt{2\pi t^3}} \exp\left[-\frac{(z - x - \mu t)^2}{2\sigma^2 t}\right], \qquad t > 0.$$

Example. Geometric Brownian motion (Example D of Section 4) is often used to model prices of assets, say shares of stock, that are traded in a perfect market. Such prices are nonnegative and usually exhibit oscillatory behavior comprised of exponential growth intermittent with exponential decay over the long run, two properties possessed by geometric Brownian motion. More importantly, if $t_0 < t_1 < \cdots < t_n$ are time points, the successive ratios

$$Y(t_1)/Y(t_0), \ldots, Y(t_n)/Y(t_{n-1}),$$

are independent random variables, so that, crudely speaking, the percentage changes over nonoverlapping time intervals are independent. Here, in a rough form, is the reasoning that supports the geometric Brownian motion as an appropriate model in a perfect market. If a ratio $Y(t + s)/Y(t)$ of a future price to a current price could be anticipated or predicted as being favorable, a number of buyers would enter the market, and their demand would tend to raise the current price $Y(t)$. Similarly, if the ratio $Y(t + s)/Y(t)$ could be predicted as being unfavorable, a number of sellers would appear and tend to depress the current price. Equilibrium obtains where the ratio cannot be predicted as being either favorable or unfavorable, that is, where price ratios over nonoverlapping time intervals are independent.

We will give an example in which the geometric Brownian motion model is used to evaluate the worth of a perpetual warrant in a stock. A warrant is an option to buy a fixed number of shares in a given stock at a stated price at any time during a specified time period. The profit to a holder of such an option is the excess of the market price over the option price. The assumption is that the holder can purchase at the stated price and resell at the market price and thus realize the potential profit.

We consider only perpetual warrants, options having no expiration dates. For such a warrant, a reasonable strategy would be to exercise the option the first time the stock price reaches some specified level we will

denote by a. By an appropriate choice of units, we may assume that the stated price in the warrant is one, so that the potential profit upon exercising the option at a market price of a is $a-1$. Of course, we need only consider $a>1$, since one would not purchase at the stated price of one if the current market price were lower.

In owning such an option, one is foregoing, in part at least, direct ownership of the stock, which is increasing at a rate of $\alpha=\mu+\frac{1}{2}\sigma^2$ per unit time, since

$$E[Y(t)|\,Y(0)=y]=y\exp\{t(\mu+\tfrac{1}{2}\sigma^2)\}.$$

One requires a higher rate of return, $\theta>\alpha$, from the option, or equivalently, discounts the potential profit of $(a-1)$ at a rate of $-\theta$ per unit time.

Let $T(a)$ be the first time the stock price reaches the level a. Then the discounted potential profit to the option holder is

$$e^{-\theta T(a)}[Y(T(a))-1]=e^{-\theta T(a)}(a-1).$$

We want to compute the expected discounted profit and then choose a to maximize this expected profit. In terms of the Brownian motion, $T(a)$ is the first time that $X(t)=\ln Y(t)$ reaches the level $\ln a$. We computed the probability density function for $T(a)$ in Theorem 5.3, and the Laplace transform in Eq. (5.5). Using (5.5) with $z=\ln a$ and $x=\ln y$, we have (see Fig. 4)

$$E[e^{-\theta T(a)}|\,Y(0)=y]=\left(\frac{y}{a}\right)^{\rho},$$

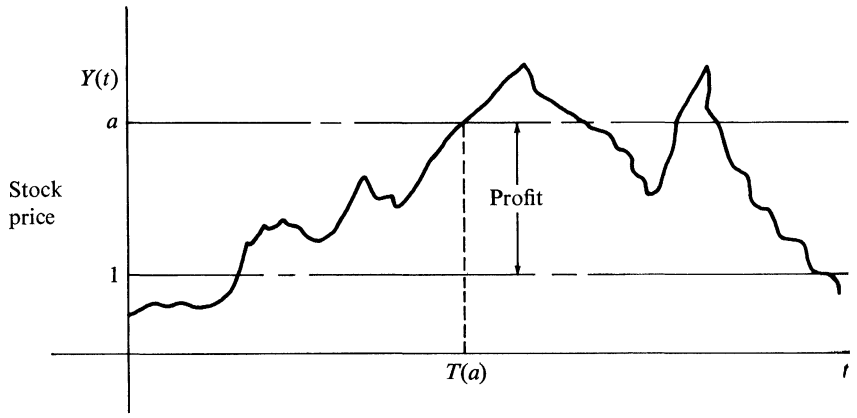


FIG. 4.

where

$$\rho = \sqrt{\frac{\mu^2}{\sigma^4} + \frac{2\theta}{\sigma^2}} - \frac{\mu}{\sigma^2}.$$

Letting $g(y, a)$ be the expected discounted profit, we have

$$\begin{aligned} g(y, a) &= (a-1)E[e^{-\theta T(a)} \mid Y(0) = y] \\ &= (a-1)\left(\frac{y}{a}\right)^{\rho}. \end{aligned}$$

We differentiate with respect to a and equate to zero to find the profit maximizing level $a = a^*$:

$$\frac{\partial g}{\partial a} = 0 = -\rho(a^* - 1)\left(\frac{y}{a^*}\right)^{\rho+1}\frac{1}{y} + \left(\frac{y}{a^*}\right)^{\rho},$$

and

$$a^* = \frac{\rho}{\rho - 1}.$$

The condition $\theta > \mu + \frac{1}{2}\sigma^2$ ensures $1 < a^* < \infty$. Given a current stock price y, the warrant has value

$$\begin{aligned} g(y, a^*) &= (a^* - 1)(y/a^*)^{\rho} \\ &= \frac{1}{\rho - 1}\left[\frac{y(\rho - 1)}{\rho}\right]^{\rho}. \end{aligned}$$

6: Multidimensional Brownian Motion

Let $\{X_1(t);\ t \geq 0\}$, ..., $\{X_N(t);\ t \geq 0\}$, be standard Brownian motion processes, statistically independent of one another in the sense that, for any finite set of time points

$$\begin{aligned} &t_{11}, t_{12}, \ldots, t_{1,n_1}, \\ &t_{21}, t_{22}, \ldots, t_{2,n_2}, \\ &\vdots \\ &t_{N,1}, t_{N,2}, \ldots, t_{N,n_N}, \end{aligned}$$

the N vectors

$$\begin{aligned} \mathbf{X}_1 &= (X_1(t_{11}), \ldots, X_1(t_{1,n_1})), \\ \mathbf{X}_2 &= (X_2(t_{21}), \ldots, X_2(t_{2,n_2})), \\ &\vdots \\ \mathbf{X}_N &= (X_N(t_{N1}), \ldots, X_N(t_{N,n_N})), \end{aligned}$$

are independent. The vector-valued process defined by

$$\mathbf{X}(t) = (X_1(t), \ldots, X_N(t)), \qquad t \geq 0,$$

is called N-dimensional Brownian motion. The motion of a particle undergoing Brownian motion in the plane and in space are described by two-dimensional and three-dimensional Brownian motions, respectively.

Consider a two-dimensional Brownian motion $\mathbf{X}(t) = (X_1(t), X_2(t))$, and let us compute the distribution of the second coordinate at the random time the first coordinate first reaches a given level $z > 0$. We let T_z be the first time t at which $X_1(t) = z$, and we then want the distribution of $Y(z) = X_2(T_z)$.

Figure 5 describes the path traced in the plane by the two-dimensional Brownian motion and displays the value $Y(z) = X_2(T_z)$.

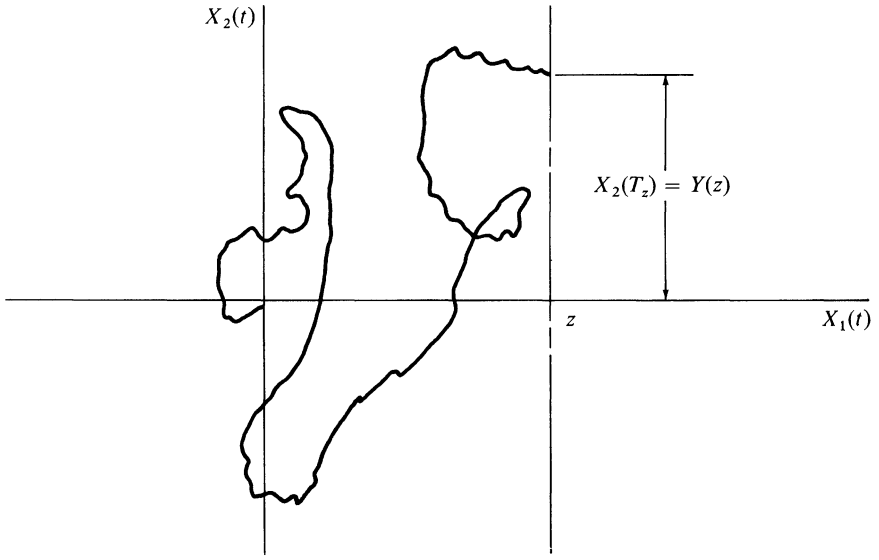


FIG. 5. Two-dimensional Brownian motion.

Fix $z > 0$. We will compute the characteristic function $\phi(u) = E[\exp\{iuY(z)\}]$. Since T is determined by the X_1 process, T and the X_2 process are independent. Hence, by the law of total probability

$$\phi(u) = \int_0^\infty E[\exp\{iuX_2(t)\}]\, dF(t),$$

where

$$F(t) = \Pr\{T_z \leq t\}$$

is the cumulative distribution function of T_z, computed explicitly just prior to Eq. (3.3). Since $X_2(t)$ is normally distributed with mean zero and variance t,

$$E[\exp\{iuX_2(t)\}] = \exp(-\tfrac{1}{2}u^2 t).$$

Clearly

$$\begin{aligned}\phi(u) &= \int_0^\infty \exp(-\tfrac{1}{2}u^2 t)\, dF(t) \\ &= E[\exp(-\tfrac{1}{2}u^2 T_z)] \\ &= \exp(-|u|z), \qquad -\infty < u < \infty.\end{aligned}$$

using (5.5), the Laplace transform for T_z, with $\theta = \frac{1}{2}u^2$, $\mu = 0$, and $\sigma^2 = 1$.

This is the characteristic function of the Cauchy probability density function

$$p(x) = \frac{1}{\pi z[1 + (x/z)^2]}, \qquad -\infty < x < \infty.$$

Even more can be said. Elementary Problem 5 is the first step in proving that the stochastic process $\{T_z;\ z \geq 0\}$ has stationary independent increments. From this it follows that

$$Y(z) = X_2(T_z) \tag{6.1}$$

also has stationary independent increments. In general, given any Markov process $\{X(t),\ t \geq 0\}$ and a process $\{T_z;\ z \geq 0\}$ having stationary independent increments and increasing sample paths, with $T_0 = 0$, it is possible to derive a new process

$$Y(z) = X[T_z], \qquad z \geq 0.$$

The process of forming Y from X is called *subordination*, and the process $\{T_z;\ z \geq 0\}$ is called the *subordinator*. Under the conditions given, $\{Y(z);\ z \geq 0\}$ will be a Markov process. If, in addition, $\{X(t);\ t \geq 0\}$ has stationary independent increments, then so will $\{Y(z);\ z \geq 0\}$.

Radial Brownian Motion.

Let $\{\mathbf{X}(t);\ t \geq 0\}$ be an N-dimensional Brownian motion process. The stochastic process defined by

$$R(t) = [X_1(t)^2 + \cdots + X_N(t)^2]^{1/2}, \qquad t \geq 0,$$

is called *Radial Brownian motion* or the *Bessel process* with parameter $\frac{1}{2}N - 1$. It is a Markov process having continuous sample paths in the

state space $[0, \infty)$. The transition probability density function from x to y is

$$p_t(x, y) = t^{-1} \exp\left\{-\frac{x^2+y^2}{2t}\right\}(xy)^{1-(N/2)} I_{(N/2)-1}\left(\frac{xy}{t}\right) y^{N-1},$$

$$t > 0, \quad x, y > 0, \tag{6.2}$$

where $I_\nu(z)$ is the modified Bessel function

$$I_\nu(z) = \sum_{k=0}^{\infty} \frac{(z/2)^{2k+\nu}}{k!\Gamma(k+\nu+1)}. \tag{6.3}$$

For $N = 1$, we use

$$I_{-1/2}(z) = \sqrt{\frac{2}{\pi z}} \cosh z$$

to get

$$p_t(x, y) = \sqrt{\frac{2}{\pi t}} \exp\left\{-\frac{x^2+y^2}{2t}\right\} \cosh\left(\frac{xy}{t}\right).$$

Comparison of this formula with (4.3) reveals that the Bessel process for $N = 1$ reduces to reflected Brownian motion.

We will investigate the case $N = 2$ shortly. When $N = 3$, the relation

$$I_{1/2}(z) = \sqrt{\frac{2}{\pi z}} \sinh z$$

produces

$$p_t(x, y) = \sqrt{\frac{2}{\pi t}} \exp\left\{-\frac{x^2+y^2}{2t}\right\} \frac{y}{x} \sinh\left(\frac{xy}{t}\right).$$

We obtain the density corresponding to $x = 0$ through continuity, letting $x \to 0$. This gives

$$p_t(0, y) = \sqrt{\frac{2}{\pi t^3}} y^2 \exp(-y^2/2t),$$

the marginal density of $R(t)$ when $N = 3$ and $R(0) = 0$.

Let us now consider the case $N = 2$, showing that the corresponding Bessel process is Markov and computing the transition density. We change to polar coordinates by defining

$$R(t) = \sqrt{X_1(t)^2 + X_2(t)^2},$$
$$\Theta(t) = \arctan[X_2(t)/X_1(t)].$$

Since Brownian motion is Markov,

$$\Pr\{R(t_n+t)\le b|\,\mathbf{X}(t_0)=\mathbf{x}_0,\ \ldots,\ \mathbf{X}(t_{n-1})=\mathbf{x}_{n-1},\ \mathbf{X}(t_n)=\mathbf{x}\}$$
$$=\Pr\{R(t_n+t)\le b|\mathbf{X}(t_n)=\mathbf{x}\}=\Pr\{R(t)\le b|\mathbf{X}(0)=\mathbf{x}\},\qquad(6.4)$$

where $0<t_0<\cdots<t_n$, $t>0$, and $\mathbf{x}_1$, ..., $\mathbf{x}_n$ are arbitrary points in the plane, $\mathbf{x}=(x_1, x_2)$. Then

$$\Pr\{R(t)\le b|\,\mathbf{X}(0)=\mathbf{x}\}$$
$$=\iint_{y_1^2+y_2^2\le b^2}\frac{1}{2\pi t}\exp\left\{-\frac{(y_1-x_1)^2+(y_2-x_2)^2}{2t}\right\}dy_1\,dy_2$$
$$=\int_0^b\left[\int_0^{2\pi}\frac{1}{2\pi t}\exp\left\{-\frac{(r\sin\theta-x_1)^2+(r\cos\theta-x_2)^2}{2t}\right\}d\theta\right]r\,dr,$$

where we have changed variables according to $y_1=r\sin\theta$, $y_2=r\cos\theta$, recalling from advanced calculus that $dy_1dy_2=r\,dr\,d\theta$. Since

$$(r\sin\theta-x_1)^2+(r\cos\theta-x_2)^2=r^2-2r(x_1\sin\theta+x_2\cos\theta)+\|\mathbf{x}\|^2$$

where $\|\mathbf{x}\|^2=x_1^2+x_2^2$,

$$\Pr\{R(t)\le b|X(0)=x\}=\int_0^b\frac{r}{2\pi t}\exp\left\{\frac{r^2+\|\mathbf{x}\|^2}{2t}\right\}I(r,\mathbf{x})\,dr$$

where

$$I(r,\mathbf{x})=\int_0^{2\pi}\exp\left\{\frac{r}{t}(x_1\sin\theta+x_2\cos\theta)\right\}d\theta.$$

Define an angle ϕ by writing

$$\sin\phi=x_1/\|\mathbf{x}\|,\qquad\cos\phi=x_2/\|\mathbf{x}\|,$$

where $\|\mathbf{x}\|=\sqrt{x_1^2+x_2^2}$. Using the trigonometric identity $\sin\phi\sin\theta+\cos\phi\cos\theta=\cos(\phi+\theta)$,

$$I(r,\mathbf{x})=\int_0^{2\pi}\exp\left\{\frac{r\|\mathbf{x}\|}{t}\cos(\phi+\theta)\right\}d\theta$$
$$=\int_0^{2\pi}\exp\left\{\frac{r\|\mathbf{x}\|}{t}\cos\theta\right\}d\theta,$$

since the integral is over the interval $\theta \in [0, 2\pi]$, over which, $\cos(\phi + \theta)$ is periodic. Now

$$\int_0^{2\pi} \exp\{\alpha \cos\theta\}\, d\theta = \int_0^{2\pi} \sum_{k=0}^{\infty} \frac{(\alpha \cos\theta)^k}{k!}\, d\theta$$
$$= \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} \int_0^{2\pi} \cos^k\theta\, d\theta.$$

But

$$\int_0^{2\pi} \cos^k\theta\, d\theta = \begin{cases} 0, & \text{if } k = 1, 3, \ldots, \\ \dfrac{k!\,2\pi}{2^k[(k/2)!]^2}, & \text{if } k = 0, 2, \ldots. \end{cases}$$

Thus, in terms of the modified Bessel function,

$$\int_0^{2\pi} \exp\{\alpha \cos\theta\}\, d\theta = 2\pi \sum_{k=0,2,\ldots} \frac{\alpha^k}{k!} \left\{\frac{\alpha^k}{2^k[(k/2)!]^2}\right\}$$
$$= 2\pi \sum_{j=0}^{\infty} \frac{\alpha^{2j}}{2^{2j}(j!)^2}$$
$$= 2\pi I_0(\alpha)$$

We have computed, then,

$$\Pr\{R(t) \le b | \mathbf{X}(0) = \mathbf{x}\} = \int_0^b \frac{r}{2\pi t} \exp\left\{\frac{r^2 + \|\mathbf{x}\|^2}{2t}\right\} 2\pi I_0\left\{\frac{r\|\mathbf{x}\|}{t}\right\} dr$$
$$= \int_0^b p_t(\|\mathbf{x}\|, r)\, dr, \tag{6.5}$$

where we take $N = 2$ in Eq. (6.2) defining p_t. We have thus shown

$$\Pr\{R(t_n + t) \le b | \mathbf{X}(t_0) = \mathbf{x}_0, \ldots, \mathbf{X}(t_n) = \mathbf{x}\} = \int_0^b p_t(\|\mathbf{x}\|, r)\, dr. \tag{6.6}$$

Let $r_j = (x_{1j}^2 + x_{2j}^2)^{1/2}$, $j = 0, \ldots, n$, where $\mathbf{x}_j = (x_{1j}, x_{2j})$. Similarly, let $\theta_j = \arctan(x_{2j}/x_{1j})$. The conditions $\mathbf{X}(t_0) = \mathbf{x}_0, \ldots, \mathbf{X}(t_n) = \mathbf{x}_n$, are equivalent to the conditions $R(t_0) = r_0, \Theta(t_0) = \theta_0, \ldots, R(t_n) = r_n, \Theta(t_n) = \theta_n$. Using this in (6.6), we will establish both the Markov property and determine the transition density function for $\{R(t), t \ge 0\}$. Beginning

with the law of total probability, where $p(\theta_0, \ldots, \theta_n)$ is the joint probability density function for $(\Theta(t_0), \ldots, \Theta(t_n))$, we have

$$\begin{aligned}
&\Pr\{R(t_n + t) \leq b | R(t_0) = r_0, \ldots, R(t_n) = r_n\} \\
&\quad = \int_0^{2\pi} \cdots \int_0^{2\pi} \Pr\{R(t_n + t) \leq b | R(t_0) = r_0, \Theta(t_0) = \theta_0, \ldots, \\
&\qquad R(t_n) = r_n, \Theta(t_n) = \theta_n\} p(\theta_0, \ldots, \theta_n)\, d\theta_0, \ldots, d\theta_n \\
&\quad = \int_0^{2\pi} \cdots \int_0^{2\pi} \left\{ \int_0^b p_t(r_n, r)\, dr \right\} p(\theta_0, \ldots, \theta_n)\, d\theta_0, \ldots, d\theta_n \\
&\quad = \int_0^b p_t(r_n, r)\, dr.
\end{aligned}$$

This verifies the Markov property and establishes (6.2) as the transition density for the two-dimensional Bessel process.

7: Brownian Paths

Considered as randomly chosen functions (as opposed to collections of random variables) the trajectories or sample paths of Brownian motion are quite remarkable. Were you asked to exhibit a continuous function that was nowhere differentiable, you might expend a considerable amount of effort to discover an example. Yet a Brownian path is "certain" (meaning the probability is one) to be such a function! This is but one example of numerous striking features of Brownian paths.

A. CONTINUITY OF PATHS

There are several ways in which a stochastic process $X(t)$, whose index set is a real interval, can be considered continuous. Three of these correspond to the different notions of *limit* for random sequences introduced in Chapter 1. We say $\{X(t)\}$ is:

(a) *Continuous in mean square* if, for every t,

$$\lim_{s \to t} E[|X(s) - X(t)|^2] = 0.$$

(b) *Continuous in probability* if, for every t and positive ε,

$$\lim_{s \to t} \Pr\{|X(s) - X(t)| > \varepsilon\} = 0.$$

(c) *Continuous almost surely* if, for every t,

$$\Pr\left\{\lim_{s\to t} X(s) = X(t)\right\} = 1.$$

The first two of these notions are decidable in terms of the finite-dimensional distributions of the process. Indeed, a process whose second moments are finite is continuous in mean square if and only if the mean value function $m(t) = E[X(t)]$ is continuous and the covariance function $\Gamma(s, t) = E[\{X(s) - m(s)\}\{X(t) - m(t)\}]$ is continuous at the diagonal $t = s$, as can be discerned readily from the expansion

$$E[|X(s) - X(t)|^2] = \Gamma(s, s) - 2\Gamma(s, t) + \Gamma(t, t) + [m(s) - m(t)]^2.$$

Chebyshev's inequality in the form

$$\Pr\{|X(s) - X(t)| > \varepsilon\} \leq \frac{1}{\varepsilon^2} E[|X(s) - X(t)|^2], \qquad \varepsilon > 0,$$

shows that every mean square continuous process is continuous in probability.

Although these are reasonable, and quite useful, concepts of continuity for many contexts, they are not adequate for all situations. Indeed, a Poisson process $N(t)$ is continuous according to all three criteria. In fact, note first that the mean value function $m(t) = \lambda t$ and covariance $\Gamma(s, t) = \lambda \min\{s, t\}$, are manifestly continuous. Moreover, for every fixed t, the event $\lim_{s\to t} X(s) \neq X(t)$ occurs only if one of the jump times of the process is at t. Since these jump times have a continuous (gamma) distribution, this event has zero probability. Thus, for every $t \geq 0$,

$$\Pr\left\{\lim_{s\to t} N(s) = N(t)\right\} = 1,$$

and a Poisson process is, according to our definition, continuous almost surely at every specified t.

But one never sees a graph of a Poisson process that shows it as a continuous function! Clearly, a more stringent criterion for continuity of random functions is called for. We say that a stochastic process $X(t)$ *almost surely has continuous paths* if, with probability one, $X(t)$ is a continuous function of t. A number of technical difficulties arise with this definition. In Chapter 1 we defined a stochastic process by specifying all its finite-dimensional distributions. Whether or not a process almost surely has continuous paths cannot be answered in terms of these distributions alone. Indeed, if $X(t)$ almost surely has continuous paths, and we define

$$\tilde{X}(t) = \begin{cases} X(t), & \text{if } t \neq \tau, \\ 0, & \text{if } t = \tau, \end{cases}$$

for some random variable τ, then, typically $\tilde{X}(t)$ does not exhibit continuous paths. Yet $X(t)$ and $\tilde{X}(t)$ have the same finite-dimensional distributions whenever τ has a continuous distribution and is independent of $\{X(t)\}$. Thus we weaken our requirement. We will say that a stochastic process, defined through its finite-dimensional distributions, almost surely has continuous paths if there is a concrete representation $\{X(t)\}$ of the process that is certain (with probability one) to be a continuous function of t.

We will now give such a representation for Brownian motion $\{X(t);\ 0 \le t \le 1\}$. Note that only time indices t in [0, 1] are considered.

The Haar functions on [0, 1] are defined by

$$H_1(t) = 1, \qquad 0 \le t \le 1,$$

$$H_2(t) = \begin{cases} 1, & 0 \le t < \frac{1}{2}, \\ -1, & \frac{1}{2} \le t \le 1, \end{cases}$$

$$H_{2^n+1}(t) = \begin{cases} 2^{n/2}, & 0 \le t < 2^{-(n+1)}, \\ -2^{n/2}, & 2^{-(n+1)} \le t \le 2^{-n}, \\ 0, & \text{otherwise.} \end{cases}$$

$$H_{2^n+j}(t) = H_{2^n+1}\left(t - \frac{j-1}{2^n}\right), \qquad j = 1, \ldots, 2^n.$$

The first six are shown in Fig. 6.

The Schauder functions are the integrals of the Haar functions, $S_k(t) = \int_0^t H_k(\tau)\, d\tau$. Their graphs are little tents, and if they are drawn, one can see

$$\max_{0 \le t \le 1} S_{2^n+j}(t) = (\tfrac{1}{2})^{(n+2)/2}, \qquad n = 0, 1, \ldots, \qquad 0 \le j \le 2^n - 1, \quad (7.1)$$

and

$$S_{2^n+j}(t) S_{2^n+k}(t) = 0, \qquad 1 \le k < j \le 2^n. \quad (7.2)$$

Now let $a(j)$, $j = 1, 2, \ldots$, be a real sequence, and set

$$b_n = \max\{|a(2^n + k)|;\ k = 1, \ldots, 2^n\}. \quad (7.3)$$

We claim the following fact. If

$$\sum_{n=0}^{\infty} b_n (\tfrac{1}{2})^{n/2} < \infty, \quad (7.4)$$

then the series

$$x(t) = \sum_{k=1}^{\infty} a(k) S_k(t)$$

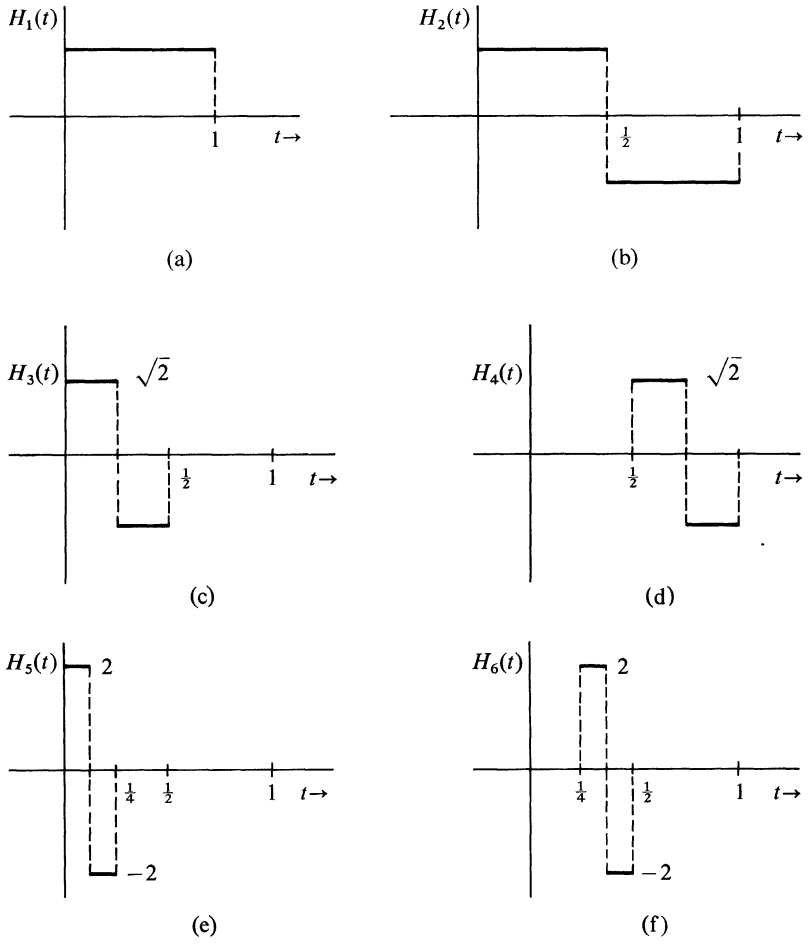


FIG. 6

converges uniformly to a continuous function of t. To validate this claim, it suffices to check the condition

$$\lim_{m,\, n\to\infty} \left| \sum_{k=m}^{m+n} a(k) \max_{0\le t\le 1} S_k(t) \right| = 0. \tag{7.5}$$

Now (7.1)–(7.3) tell us

$$\left| \sum_{k=2^j+1}^{2^{j+1}} a(k) \max_{0\le t\le 1} S_k(t) \right| \le b_j 2^{-(j+2)/2}, \tag{7.6}$$

so by grouping the summands in (7.5) according to indices $k = 2^j$, $2^j + 1, \ldots, 2^{j+1}$, we see that whenever $m > 2^N$, the Cauchy sum in (7.5) is smaller than $\sum_{j=N}^{\infty} b_j 2^{-(j+2)/2}$, which converges to zero under (7.4). Then $x(t)$ is a continuous function of t whenever the coefficients satisfy (7.4), as claimed.

Now let $A_1, A_2, A_3, \ldots$ be independent normally distributed random variables having zero means and unit variances. Let

$$B_n = \max\{|A_k| : 2^n < k \leq 2^{n+1}\}. \tag{7.7}$$

Then

$$X(t) = \sum_{k=1}^{\infty} A_k S_k(t) \tag{7.8}$$

is a continuous function of t whenever $\sum_{n=0}^{\infty} B_n(\frac{1}{2})^{n/2} < \infty$. We claim

$$\Pr\left\{\sum_{n=0}^{\infty} B_n(\tfrac{1}{2})^{n/2} < \infty\right\} = 1,$$

so that, with "certainty," (7.8) defines a continuous function. To validate this claim, we first implement integration by parts on the normal integral to get

$$\begin{aligned}\Pr\{|A_k| \geq x\} &= \frac{2}{\sqrt{2\pi}} \int_x^{\infty} \exp(-u^2/2)\, du \\ &= \frac{2}{\sqrt{2\pi}} \left\{\frac{\exp(-x^2/2)}{x} - \int_x^{\infty} \frac{\exp(-u^2/2)}{u^2}\, du\right\} \\ &\leq \frac{2}{\sqrt{2\pi}} \frac{\exp(-x^2/2)}{x}.\end{aligned}$$

Thus

$$\sum_{n=2}^{\infty} \Pr\{|A_n| > 2\sqrt{\log n}\} \leq K \sum_{n=2}^{\infty} \frac{1}{n^2\sqrt{\log n}},$$

where K is a constant. The sum on the right converges, so the Borel–Cantelli lemma (Chapter 1) implies that only finitely many values of $|A_n|$ exceed $2\sqrt{\log n}$. This means, that only finitely many values of $B_j = \max\{|A_n| : 2^j < n \leq 2^{j+1}\}$ exceed $2\sqrt{\log 2}\sqrt{j}$. Since $\sum \sqrt{j}(\frac{1}{2})^{j/2} < \infty$, this verifies the convergence, with probability one, of $\sum B_n(\frac{1}{2})^{n/2}$, and consequently the continuity of $X(t) = \sum A_k S_k(t)$.

We have yet to show that $X(t)$ is Brownian motion. If every finite-dimensional vector $\{X(t_1), \ldots, X(t_k)\}$ of a process has a multivariate normal distribution, we call the process *Gaussian*. A Gaussian process is determined by its mean value and covariance functions, because these parameters uniquely specify all the finite-dimensional multivariate normal distributions. Thus, to complete our endeavor, we need only show that: (1) $X(t)$ is Gaussian; (2) $E[X(t)] = 0$; and (3) $E[X(t)X(s)] = \min\{s, t\}$.

The first two properties are easy to check. Each partial sum $X_n(t) = \sum_{k=0}^{n} A_k S_k(t)$ is Gaussian (why?), and this property is preserved in the limit. Indeed, it is easy to ascertain that $X_n(t)$ converges to $X(t)$ in mean square, which will justify the interchange of limit and expectation in

$$E[X(t)] = \lim_{n\to\infty} \sum_{k=1}^{n} E[A_k]S_k(t) = 0.$$

All that remains is to determine the covariance between $X(t)$ and $X(s)$ and see if it is identical with the covariance of standard Brownian motion, namely $\min\{s, t\}$. That is, we wish to verify the equation

$$\begin{aligned}\min\{s, t\} &\stackrel{?}{=} E[X(s)X(t)] \\ &= \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} E[A_j A_k]S_j(s)S_k(t) \\ &= \sum_{k=1}^{\infty} S_k(s)S_k(t).\end{aligned}$$

This is a question purely in classical analysis. But to keep our treatment self-contained, and to demonstrate the power of our methods and the content of our theorems, we will use a martingale argument. In Problem 24 of Chapter 6, we asserted

$$\sum_{k=1}^{n} a_k H_k(Z) = E[f(Z)|\, Y_1, \ldots, Y_n],$$

where $f(\tau)$ was an arbitrary function satisfying

$$\int_0^1 |f(\tau)|\, d\tau < \infty, \qquad a_k = \int_0^1 f(\tau)H_k(\tau)\, d\tau, \qquad \text{and} \qquad Y_k = H_k(Z),$$

for Z uniformly distributed on [0, 1]. Thus these partial sums form a Doob's martingale that converges. If $\int_0^1 |f(\tau)|^2\, d\tau < \infty$, the martingale is actually square integrable and

$$\sum_{k=1}^{n} a_k H_k(Z) \to E[f(Z)|\, Y_1, Y_2, \ldots] = f(Z),$$

the convergence occurring in mean square as $n \to \infty$. The last equality is a consequence of Z being determined by the infinite sequence $Y_1, Y_2, \ldots$. Evaluate the mean square to see

$$\int_0^1 |f(\tau) - \sum_{k=1}^n a_k H_k(\tau)|^2 \, d\tau = \int_0^1 \{f(\tau)\}^2 \, d\tau - 2\sum_{k=1}^n a_k \int_0^1 f(\tau)H_k(\tau) \, d\tau + \int_0^1 \left\{\sum_{k=1}^n a_k H_k(\tau)\right\}^2 d\tau = \int_0^1 \{f(\tau)\}^2 \, d\tau - \sum_{k=1}^n a_k^2.$$

Since this converges to zero as $n \to \infty$, we deduce $\int_0^1 \{f(\tau)\}^2 \, d\tau = \sum_{k=1}^\infty a_k^2$. Applying this formula first to $[f(\tau) + g(\tau)]$ and then to $f(\tau)$ and $g(\tau)$ and subtracting leads to the (so-called) Parseval relation

$$\int_0^1 f(\tau)g(\tau) \, d\tau = \sum_{k=1}^\infty a_k b_k, \tag{7.9}$$

where $b_k = \int_0^1 g(\tau)H_k(\tau) \, d\tau$. Fix $s < t$ and let

$$f(\tau) = \begin{cases} 1, & 0 \le \tau \le s, \\ 0, & s < \tau \le 1, \end{cases}$$

and

$$g(\tau) = \begin{cases} 1, & 0 \le \tau \le t, \\ 0, & t < \tau \le 1. \end{cases}$$

Then $a_k = \int_0^1 f(\tau)H_k(\tau) \, d\tau = S_k(s)$, and $b_k = S_k(t)$, while $\int_0^1 f(\tau)g(\tau) \, d\tau = s$ whenever $s < t$. Substitution into (7.9) gives $s = \sum_{k=0}^\infty S_k(s)S_k(t)$. The same argument works when $t < s$ to give our desired

$$\min\{s, t\} = \sum_{k=1}^\infty S_k(s)S_k(t).$$

This completes our proof that $X(t) = \sum_{k=1}^\infty A_k S_k(t)$ is a Brownian motion process whose paths are continuous functions of t with probability one.

If a Brownian motion $B(t)$ whose index set is $0 \le t < \infty$ is desired, first set $W(t) = tX(1/t)$, for $1 \le t < \infty$, and then set $B(t) = W(1 + t) - W(1)$, $t \ge 0$. We leave it to the reader to check that this is indeed the desired process. (It is Gaussian and has the required mean and covariance functions.)

A final curiosity that emanates from the construction: Array the digits in the decimal expansion of a uniform (0, 1) random variable $U = Z_1 Z_2 Z_3 \cdots$ diagonally in an infinite matrix as shown:

$$\begin{array}{ccccccc} Z_1 & & Z_3 & & Z_6 & & Z_{10} \quad \cdots \\ & Z_2 & & Z_5 & & Z_9 \quad \cdots & \\ & & Z_4 & & Z_8 \quad \cdots & & \\ & & & Z_7 \quad \cdots & & & \end{array}$$

The rows give the decimal expansion for

$$\begin{aligned} U_1 &= \cdot\, Z_1 Z_3 Z_6 Z_{10} \cdots, \\ U_2 &= \cdot\, Z_2 Z_5 Z_9 Z_{14} \cdots, \\ U_3 &= \cdot\, Z_4 Z_8 Z_{13} Z_{19} \cdots, \end{aligned}$$

which are independent and also uniformly distributed over (0, 1). Let Φ^{-1} be the inverse function to the normal integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt,$$

and set $A_k = \Phi^{-1}(U_k)$. These random variables are independent and, being inverse probability transforms, are normally distributed with zero means and unit variances. Use them in formula (7.8) to construct a Brownian motion process. We have then formed an entire Brownian motion process using as our sole source of "randomness" a single uniform random variable U! In this sense, Brownian motion has no more "randomness" than the experiment of drawing a single number by chance!

B. THE SQUARED VARIATION

Let $X(t)$ be a standard Brownian motion. We will not show that $X(t)$ is nowhere differentiable, although, as mentioned earlier, this is indeed true. What we will do, however, will lend support to this conclusion. We will establish that, for every fixed $t > 0$,

$$\lim_{n \to \infty} \sum_{k=1}^{2^n} \left[X\left(\frac{k}{2^n} t\right) - X\left(\frac{k-1}{2^n} t\right) \right]^2 = t. \tag{7.10}$$

This convergence takes place both in mean square and with probability one, or almost surely.

An elementary calculus approach to the limit on the left suggests the formula

$$\int_0^t [dX(\tau)]^2 = t = \int_0^t d\tau.$$

The same formula expressed in differentials reads $[dX(t)]^2 = dt$! A typical feeling on seeing this for the first time is disbelief accompanied by a strong desire to check the analysis carefully and preclude the possibility of error. No error has been made, as we invite you to check for yourself shortly. In fact, the differential formula $dX(t)^2 = dt$ can be endowed with a precise meaning in which it is not only true, but highly useful, but this development is deferred to Chapter 15 on diffusion processes.

Before proceeding to the proof, let us draw an easy corollary of (7.10):

$$\lim_{n\to\infty} \sum_{k=1}^{2^n} \left| X\left(\frac{k}{2^n}t\right) - X\left(\frac{k-1}{2^n}t\right)\right| = \infty. \tag{7.11}$$

In words, the total variation of a Brownian path is infinite (with probability one). This suggests, but does not imply, the nondifferentiable nature of the paths mentioned earlier. The infinite total variation results from the inequality

$$\sum_{k=1}^{2^n} \left| X\left(\frac{k}{2^n}t\right) - X\left(\frac{k-1}{2^n}t\right)\right| \geq \frac{\displaystyle\sum_{k=1}^{2^n} \left[X\left(\frac{k}{2^n}t\right) - X\left(\frac{k-1}{2^n}t\right)\right]^2}{\displaystyle\max_{j=1,\ldots,2^n} \left|X\left(\frac{j}{2^n}t\right) - X\left(\frac{j-1}{2^n}t\right)\right|}.$$

The numerator on the right converges to t, while the denominator vanishes because Brownian paths are continuous and thus uniformly continuous over bounded intervals. Thus the left hand side must become infinite which validates (7.11).

The proof of the squared variation formula is greatly eased if we introduce some briefer notation. Let

$$\Delta_{nk} = X\left(\frac{k}{2^n}t\right) - X\left(\frac{k-1}{2^n}t\right), \qquad k = 1, \ldots, 2^n,$$

and

$$W_{nk} = \Delta_{nk}^2 - t/2^n, \qquad k = 1, \ldots, 2^n.$$

We wish to show that $\sum_{k=1}^{2^n} \Delta_{nk}^2 \to t$, or what is the same, $\sum_{k=1}^{2^n} W_{nk} \to 0$. For each n, the random variables in $\{W_{nk}\}_{k=1}^{2^n}$ are independent, identically distributed, and

$$E[W_{nk}] = E[\Delta_{nk}^2] - t/2^n = 0, \qquad E[W_{nk}^2] = 2t^2/4^n.$$

The last computation is the fourth moment of the normally distributed Δ_{nk}. In fact, if Δ is normally distributed with mean zero and variance σ^2, then $E[\Delta^{2m}] = 1 \cdot 3 \cdots (2m-1)\sigma^{2m}$, which is readily confirmed inductively by differentiating the normal characteristic function. Then, $E[W_{kn} W_{jn}] = 0$ if $j \neq k$, and squaring the sum leads to

$$E\left[\left\{\sum_{k=1}^{2^n} W_{kn}\right\}^2\right] = \sum_{k=1}^{2^n} E[W_{nk}^2] = 2^{n+1}t^2/4^n = 2t^2/2^n.$$

Since $2t^2/2^n \to 0$ as $n \to \infty$, this immediately shows our desired squared variation formula holds when the limit is understood in the mean square sense. To get convergence with probability one let $\varepsilon > 0$ be given and apply Chebyshev's inequality to see

$$\Pr\left\{\left|\sum_{k=1}^{2^n} W_{nk}\right| > \varepsilon\right\} \leq \frac{2t^2}{\varepsilon^2}\left(\frac{1}{2}\right)^n.$$

Since $\sum (\frac{1}{2})^n < \infty$, the Borel–Cantelli lemma implies that $\left|\sum_{k=1}^{2^n} W_{nk}\right| > \varepsilon$ can occur for only finitely many values of n. Since $\varepsilon > 0$ is arbitrary, we must have (with probability one)

$$\lim_{n\to\infty} \sum_{k=1}^{2^n} W_{nk} = 0.$$

This is equivalent to the squared variation relation (7.10).

While $X(t)$ is not differentiable, in Section 8 of Chapter 9 we will attach a meaning to expressions like $\int_0^t f(\tau)\, dX(\tau)$ by defining the integral to be the mean square limit of the approximating sums. A more general stochastic integral will be developed in Chapter 15.

C. THE LAW OF THE ITERATED LOGARITHM

The principal form of the celebrated law of the iterated logarithm for Brownian motion $X(t)$ states that

$$\limsup_{t\downarrow 0} \frac{X(t)}{\sqrt{2t \log\log(1/t)}} = 1 \tag{7.12}$$

is a certain event or, equivalently, is an event of probability one. There are many variations and generalizations. For example, since $tX(1/t)$ is also a Brownian motion, we have

$$\limsup_{t\downarrow 0} \frac{tX(1/t)}{\sqrt{2t \log\log(1/t)}} = 1,$$

or, with $s = 1/t$,

$$\limsup_{s\uparrow\infty} \frac{X(s)}{\sqrt{2s \log\log s}} = 1.$$

Given a positive ε, on the one hand this means that no matter how large a value s we take, there are values $t > s$ for which

$$\frac{1}{\sqrt{t}} X(t) > (1 - \varepsilon)\sqrt{2 \log\log t},$$

while on the other hand, we may guarantee for each sample path

$$\frac{1}{\sqrt{t}} X(t) < (1 + \varepsilon)\sqrt{2 \log\log t}, \qquad \text{for all} \quad t > s,$$

by choosing s sufficiently large. In this form, we see that the law of the iterated logarithm furnishes a remarkable answer to what is basically a simple question. For any fixed t, $X(t)/\sqrt{t}$ is normally distributed with mean zero and variance one, and so

$$\Pr\left\{\frac{1}{\sqrt{t}} X(t) > K\sqrt{2 \log\log t}\right\} = 1 - \Phi(K\sqrt{2 \log\log t}),$$

where $\Phi(x)$ is the normal integral. For $K = 1$ and $t = 10^{10}$, this probability is quite small, approximately 0.006. On the other hand, it is conceivable that the probability of

$$\frac{1}{\sqrt{t}} X(t) > K\sqrt{2 \log\log t},$$

for *some* $t > 10^{10}$ could be much larger. The law of the iterated logarithm says "no" if $K > 1$ but "yes," with the probability approaching one, if $K < 1$.

We content ourselves with proving only half of the law, namely

$$\limsup_{t\downarrow 0} \frac{X(t)}{\sqrt{2t \log\log(1/t)}} \leq 1, \tag{7.13}$$

with probability one. We begin by applying the maximal inequality to the nonnegative martingale

$$Z(s) = \exp\{\alpha X(s) - \tfrac{1}{2}\alpha^2 s\}, \qquad \alpha > 0,$$

to deduce, for $\alpha > 0$, $\beta > 0$,

$$\begin{aligned}\Pr\left\{\max_{k=1,\ldots,2^n} \exp\left\{\alpha X\left(\frac{k}{2^n}t\right) - \frac{1}{2}\alpha^2\frac{k}{2^n}t\right\} > e^{\alpha\beta}\right\} &\le e^{-\alpha\beta}E[Z(t)] \\ &= e^{-\alpha\beta}E[Z(0)] \\ &= e^{-\alpha\beta}.\end{aligned}$$

This holds for every $n = 1, 2, \ldots$, and since the paths of $X(s)$ are continuous, we may let $n \to \infty$ to see

$$\begin{aligned}e^{-\alpha\beta} &\ge \Pr\left\{\sup_{0\le s\le t} \exp\{\alpha X(s) - \tfrac{1}{2}\alpha^2 s\} > e^{\alpha\beta}\right\} \\ &= \Pr\left\{\sup_{0\le s\le t} \{X(s) - \tfrac{1}{2}\alpha s\} > \beta\right\}, \qquad \alpha > 0.\end{aligned}$$

Fix a value θ, with $0 < \theta < 1$, set

$$h(t) = \sqrt{2t \log\log(1/t)},$$

and for $\varepsilon > 0$, choose

$$\alpha = \alpha_n = (1 + 2\varepsilon)\theta^{-n}h(\theta^n), \qquad \beta = \beta_n = \tfrac{1}{2}h(\theta^n).$$

We get

$$\alpha\beta = (1 + 2\varepsilon)\log\log\theta^{-n} = (1 + 2\varepsilon)\log nc, \qquad \text{for} \quad c = \log(1/\theta) > 0.$$

Thus $e^{-\alpha\beta} = (nc)^{-(1+2\varepsilon)}$. Since $\sum_{n=1}^{\infty}(nc)^{-(1+2\varepsilon)} < \infty$, the Borel–Cantelli lemma applies, and we conclude that

$$\sup_{0\le s\le t} \{X(s) - \tfrac{1}{2}\alpha_n s\} \le \beta_n, \tag{7.14}$$

holds for all but finitely many values of n. In particular, we may assume there is some integer N, random since it depends on the particular path being studied, but for which (7.14) holds whenever $n > N$. If $t < \theta^N$, then when covered by the interval $\theta^n < t \le \theta^{n-1}$, we have $n > N$, so that

$$\begin{aligned}\beta_n &\ge \sup_{0\le s\le t} \{X(s) - \tfrac{1}{2}\alpha_n s\} \\ &\ge X(t) - \tfrac{1}{2}\alpha_n t,\end{aligned}$$

and

$$\begin{aligned}X(t) &\le \tfrac{1}{2}\alpha_n t + \beta_n \\ &\le \tfrac{1}{2}\alpha_n \theta^{n-1} + \beta_n \\ &= \tfrac{1}{2}(1 + 2\varepsilon)\theta^{-1}h(\theta^n) + \tfrac{1}{2}h(\theta^n) \\ &= \frac{1}{2}\left\{\frac{1 + 2\varepsilon}{\theta} + 1\right\}h(\theta^n).\end{aligned}$$

Since $h(t)$ is an increasing function for t near zero, $h(\theta^n) \leq h(t)$, and

$$X(t) \leq \frac{1}{2}\left\{\frac{1+2\varepsilon}{\theta}+1\right\}h(t).$$

This inequality holds for all t sufficiently near the origin. To be precise, it holds for all $t \leq \theta^N$. Thus, with probability one,

$$\limsup_{t\downarrow 0}\frac{X(t)}{h(t)} \leq \frac{1}{2}\left\{\frac{1+2\varepsilon}{\theta}+1\right\}.$$

Since we have placed no restrictions on θ other than $0<\theta<1$, let $\theta \to 1$ to conclude

$$\limsup_{t\downarrow 0}\frac{X(t)}{h(t)} \leq 1+\varepsilon,$$

with probability one.

Since ε is an arbitrary positive number, we have proved the half of the law of the iterated logarithm that is stated in (7.13).

Elementary Problems

In these problems, $X(t)$ is standard Brownian motion.

1. Let T_0 be the largest zero of $X(\tau)$ not exceeding t. Establish the formula

$$\Pr\{T_0 < t_0\} = \frac{2}{\pi}\arcsin\sqrt{t_0/t}.$$

Hint: Use Theorem 3.1.

2. Let T_1 be the smallest zero of $X(\tau)$ exceeding t. Show that:

(a) $\Pr\{T_1 < t_1\} = \dfrac{2}{\pi}\arccos\sqrt{t/t_1}$.

(b) $\Pr\{T_0 < t_0,\ T_1 > t_1\} = \dfrac{2}{\pi}\arcsin\sqrt{t_0/t_1}$.

3. Verify that $E(X(t)X(s)|X(0)=0) = \min(t, s)$.

4. Show that the density

$$p(t, x, y) = \frac{1}{\sqrt{2\pi t}}\exp[-(x-y)^2/2t]$$

satisfies the heat equation

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2}.$$

5. Let $T(\lambda)$ be the first passage time for reaching $\lambda > 0$ when $X(0) = 0$. Prove that the distribution of $T(\lambda_1 + \lambda_2)$ is the same as the distribution of the sum of $T(\lambda_1)$ and $T(\lambda_2)$, where $T(\lambda_1)$ and $T(\lambda_2)$ are regarded as independent random variables, $\lambda_1, \lambda_2 > 0$.

Hint: Verify $\phi_{\lambda_1+\lambda_2}(\theta) = \phi_{\lambda_1}(\theta)\phi_{\lambda_2}(\theta)$, where $\phi_\lambda(\theta)$ is the Laplace transform of $T(\lambda)$, given in Eq. (5.3).

6. Determine the covariance functions for

$$U(t) = e^{-t}X(e^{2t}), \qquad t \geq 0,$$

and

$$V(t) = X(t) - tX(1), \qquad 0 \leq t \leq 1.$$

Solution:

$$E[U(t)U(s)] = \exp\{-|t-s|\},$$

and

$$E[V(t)V(s)] = t(1-s), \qquad \text{for} \quad t \leq s.$$

7. For a standard Brownian motion $X(t)$ and constants $\alpha > 0$, $\beta > 0$, establish

$$\Pr\{X(t) \leq \alpha t + \beta \quad \text{for all} \quad t \geq 0 | X(0) = w\} = 1 - e^{-2\alpha(\beta - w)}, \qquad \text{for} \quad w \leq \beta.$$

Hint: Apply Corollary 5.1 to the Brownian motion with drift $W(t) = X(t) - \alpha t - w$.

8. Let $W = \int_0^t X(s)\,ds$. Verify that $E[W] = 0$ and $E[W^2] = t^3/3$.

Hint: Validate and complete the computation

$$\begin{aligned} E[W^2] &= E\left[\left\{\int_0^t X(s)\,ds\right\}^2\right] \\ &= E\left[\left\{\int_0^t X(u)\,du\right\}\left\{\int_0^t X(v)\,dv\right\}\right] \\ &= \int_0^t\int_0^t E[X(u)X(v)]\,du\,dv \\ &= 2\int_0^t \left\{\int_0^v u\,du\right\} dv. \end{aligned}$$

9. Derive the conditional distribution of $W = \int_0^t X(s)\,ds$ given that $X(t) = x$.

Hint: W and $X(t)$ have a joint normal distribution.

Solution: Given $X(t) = x$, W is normally distributed with mean $E[W|X(t) = x] = \frac{1}{2}tx$ and variance $E[(W - \frac{1}{2}tx)^2 | X(t) = x] = t^3/12$.

10. Let T be the first time the Brownian motion process crosses the line $l(t) = \alpha + \beta t$, $(\alpha > 0, \beta > 0)$. Determine the Laplace transform of T.

Hint: Use the second martingale of (5.1) yielding

$$E[\exp\{\lambda(\alpha + \beta T) - \tfrac{1}{2}\lambda^2 T\}] = 1$$

and then a change of variable $\lambda\beta - \frac{1}{2}\lambda^2 = z$.

11. Let $Y(t) = e^{X(t)}$ be geometric Brownian motion. Determine the diffusion coefficients

$$\lim_{h \downarrow 0} \frac{E[Y(t+h) - Y(t) | Y(t) = y]}{h} = b(y), \qquad 0 < y < \infty,$$

and

$$\lim_{h \downarrow 0} \frac{E[\{Y(t+h) - Y(t)\}^2 | Y(t) = y]}{h} = a(y), \qquad 0 < y < \infty.$$

12. Use relation (5.5) to evaluate the integrals

$$\int_0^\infty \frac{1}{\sqrt{t}} \exp\left\{-\left(at + \frac{b}{t}\right)\right\} dt, \qquad a, b > 0;$$

$$\int_0^\infty \frac{1}{t^{3/2}} \exp\left\{-\left(at + \frac{b}{t}\right)\right\} dt.$$

13. Prove that $\Pr\{M(t) > \xi | X(t) = M(t)\} = \exp(-\xi^2/2t)$, where $M(t) = \max_{0 \le u \le t} X(u)$.

Hint: Let $Y(t) = M(t) - X(t)$. Find the conditional distribution of $M(t)$ given $Y(t) = 0$.

14. Validate the identities

(i) $$E[\exp\{\lambda \int_0^t X(s)\, ds\}] = \exp(\lambda^2 t^3/6), \qquad -\infty < \lambda < \infty,$$

and

(ii) $$E[\exp\{\lambda \int_0^t sX(s)\, ds\}] = \exp(\lambda^2 t^5/15), \qquad -\infty < \lambda < \infty.$$

15. Let $R(t) = [X_1(t)^2 + \cdots + X_m(t)^2]^{1/2}$ be the radial Brownian motion or Bessel process in m dimensions. (a) Validate that $R(t)^2 - mt$ is a martingale. (b) Use the martingale optional sampling theorem to establish that $E[T] = r^2/m$, where $T = \inf\{t \ge 0;\ R(t) \ge r\}$ is the first time the m-dimensional Brownian motion $[X_1(t), \ldots, X_m(t)]$ reaches a distance r from the origin.

Problems

We use the notation

$$M(t) = \max_{0 \le u \le t} X(u),$$

and

$$Y(t) = M(t) - X(t),$$

where $X(t)$ is standard Brownian motion.

1. Prove that $Y(t) = M(t) - X(t)$ is a continuous-time Markov process.

Hint: Note that for $t' < t$,

$$Y(t) = \max\{\max_{t' \le u \le t} (X(u) - X(t')),\ Y(t')\} - (X(t) - X(t')).$$

2. Show that the stochastic process $Y(t) = M(t) - X(t)$ and the stochastic process $|X(t)|$ are equivalent. (Two processes are said to be equivalent if the finite-dimensional distributions are the same.)

Hint: Since $|X(t)|$ and $Y(t)$ are both Markov processes, it is enough to prove that the density functions of

$$\Pr\{Y(t) < y \,|\, Y(t_0) = y_0,\ t_0 < t\} \qquad \text{and} \qquad \Pr\{|X(t)| < y \,|\, |X(t_0)| = y_0,\ t_0 < t\}$$

are identical.

To compute the left-hand side, use the representation of $Y(t)$ in Problem 1.

3. Prove that the probability of at least one zero of $Y(t)$ in the interval (t_0, t_1) is $(2/\pi) \arccos\sqrt{t_0/t_1}$.

4. Let $T_1^*(T_0^*)$ be the smallest (largest) zero of $Y(\tau) = M(\tau) - X(\tau)$ exceeding (not exceeding) t. Show that T_0^* and T_1^* possess the same distribution as T_0 and T_1, respectively, as defined in Elementary Problems 1 and 2.

5. For $a \cdot b > 0$, prove

$$\Pr\{X(\tau) \text{ is not zero in } (0, t) | X(0) = a,\ X(t) = b\} = 1 - e^{-2ab/t}.$$

Hint: Use the function $A_t(x, y)$ of (3.9).

6. Prove that, for $\alpha, \beta > 0$,

$$\Pr\{X(u) < \alpha u + \beta,\ 0 \le u \le 1 | X(0) = X(1) = 0\} = 1 - e^{-2\beta(\beta+\alpha)}.$$

Hint: Use Theorem 2.1 to establish the identity

$$\begin{aligned}\Pr\{X(u) < \alpha u + \beta,\ 0 \le u \le 1 | X(0) = X(1) = 0\} \\ = \Pr\{X(u) < 0,\ 0 \le u \le 1 | X(0) = -\beta,\ X(1) = -\beta - \alpha\},\end{aligned}$$

and then consult Problem 5.

7. Find the conditional probability that $X(t)$ is not zero in the interval (t_0, t_2), given that it is not zero in the interval (t_0, t_1), $0 < t_0 \leq t_1 \leq t_2$.

Answer:

$$\frac{\arcsin\sqrt{t_0/t_2}}{\arcsin\sqrt{t_0/t_1}}.$$

8. Show that the probability that $X(t)$ is not zero in $(0, t_2)$, given that it is not zero in the interval $(0, t_1)$, $0 < t_1 < t_2$, is $\sqrt{t_1/t_2}$.

Hint: Compute

$$\Pr\{X(t) \neq 0,\ 0 < t_0 \leq t \leq t_2 | X(t) \neq 0,\ 0 < t_0 \leq t \leq t_1\},$$

and then let $t_0 \to 0$.

9. Show that the probability of the event $|X(t_1) - X(t_0)| > \xi$, given that $X(t)$ takes on an extreme value [$X(t)$ has two extreme values] over the interval (t_0, t_1) at either t_0 or t_1, is $\exp(-\xi^2/2(t_1 - t_0))$, $t_0 > 0$.

Hint: Prove the following statements:

(i) The event of the problem can take place in any one of four ways: (A) $X(t_0)$ is a minimum, (B) $X(t_0)$ is a maximum, (C) $X(t_1)$ is a minimum, (D) $X(t_1)$ is a maximum.

(ii) The conditional probability, given any one of (A), (B), (C), (D) that any other one of (A), (B), (C), (D) occurs is zero.

(iii)

$$\begin{aligned}\Pr\{|X(t_1) - X(t_0)| > \xi | &\text{(A), (B), (C), or (D) occurs}\} \\ &= \sum_{\alpha = \text{(A), (B), (C), (D)}} \Pr\{|X(t_1) - X(t_0)| > \xi | \alpha\} \\ &\quad \times \Pr\{\alpha | \text{(A), (B), (C), or (D) occurs}\} \\ &= \exp[-\xi^2/2(t_1 - t_0)].\end{aligned}$$

(Use Elementary Problem 13 and the reflection principle.)

10. Prove that

$$\begin{aligned}&\Pr\{X(\tau) \neq 0 \quad \text{in} \quad 0 < t < \tau < u < 1 | X(0) = X(1) = 0\} \\ &\quad = \frac{2}{\pi} \arccos \sqrt{\frac{u - t}{u(1 - t)}}.\end{aligned}$$

Hint: Compute the quantity

$$\begin{aligned}2 \int_{\alpha=0}^{\infty} \int_{\tau=u}^{1} &\Pr\{X(t) = \alpha,\ T(\alpha) = \tau - t,\ X(1 - \tau) = 0 | X(0) = 0\} \\ &\times [\Pr\{X(1) = 0 | X(0) = 0\}]^{-1}\, d\alpha\, d\tau,\end{aligned}$$

where $T(\alpha)$ denotes the time at which the Brownian particle first becomes 0 starting from $X(0) = \alpha$ [see (3.7)], and

$$\begin{aligned}\frac{d}{du}\left[\frac{2}{\pi}\arccos\sqrt{\frac{u-t}{u(1-t)}}\right] &= -\frac{2}{\pi}\frac{1}{\sqrt{1-(u-t)/u(1-t)}}\left(\frac{1}{\sqrt{1-t}}\right)\frac{t}{2\sqrt{1-(t/u)u^2}} \\ &= -\frac{\sqrt{t}}{\pi}\frac{1}{(\sqrt{1-u})(\sqrt{u-t})u},\end{aligned}$$

which proves the result.

11. Establish the identity

$$E\left[\exp\left\{\lambda\int_0^t f(s)\,X(s)\,ds\right\}\right] = \exp\left\{\lambda^2\int_0^t f(v)\left[\int_0^v uf(u)\,du\right]dv\right\}, \qquad -\infty < \lambda < \infty$$

for any continuous function $f(s)$, $0 \le s < \infty$.

12. Prove that $\Pr\{X(1) \le x | X(u) \ge 0,\ 0 \le u \le 1\} = 1 - \exp(-x^2/2)$.

Hint: $X'(t) = X(1) - X(1-t)$ is also a Brownian motion. The desired probability, in terms of $X'(t)$, is

$$\Pr\{X'(1) \le x | M'(1) = X'(1)\},$$

where $M'(t) = \max_{0 \le x \le t} X'(x)$. Now consult Elementary Problem 13.

13. For $a > 0$, $b < a$, show

$$\Pr\left\{\sup_{t \ge 0}\frac{b + X(t)}{1+t} \ge a\right\} = e^{-2a(a-b)}.$$

Then show that the left-hand side, hence also the right, equals

$$\Pr\left\{\sup_{0 \le u \le 1} X(u) \ge a | X(1) = b\right\}.$$

14. Prove Kolmogorov's inequality for Brownian motion:

$$\Pr\left\{\sup_{0 \le u \le t} |X(u)| > \varepsilon\right\} \le t/\varepsilon^2, \qquad \varepsilon > 0.$$

15. (Continuation). Use Kolmogorov's inequality to show

$$\lim_{t \to \infty}\frac{1}{t}X(t) = 0.$$

Hint: Set $\varepsilon = 2^{2n/3}$ and $t = 2^n$, and apply the Borel–Cantelli lemma.

16. For $n = 1, 2, \ldots$ and $k = 1, \ldots, 2^n$, set

$$\Delta_{nk} = X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)$$

where $X(t)$ is standard Brownian motion.

Show $E[S_{n+1}|S_n] = \frac{1}{2}(S_n + 1)$, where $S_n = \sum_{k=1}^{2^n} \Delta_{nk}^2$.

Hint: Use Theorem 2.1 to establish

$$E[\Delta_{n+1,\,2k-1}^2 + \Delta_{n+1,\,2k}^2 | X(j/2^n), \quad j = 1, \ldots, 2^n] = \tfrac{1}{2}(\Delta_{nk}^2 + 1).$$

Then sum both sides.

17. Using the notation of Problem 16, show $E[S_n|S_{n+1}] = S_{n+1}$.

Hint: Use symmetry to argue

$$\begin{aligned} &E[\Delta_{nk}^2 | \Delta_{n+1,\,2k-1}^2, \quad \Delta_{n+1,\,2k}^2] \\ &\quad = E[(\Delta_{n+1,\,2k-1} + \Delta_{n+1,\,2k})^2 | \Delta_{n+1,\,2k-1}^2, \quad \Delta_{n+1,\,2k}^2] \\ &\quad = \tfrac{1}{2}(\Delta_{n+1,\,2k-1} + \Delta_{n+1,\,2k})^2 + \tfrac{1}{2}(\Delta_{n+1,\,2k-1} - \Delta_{n+1,\,2k})^2 \\ &\quad = \Delta_{n+1,\,2k-1}^2 + \Delta_{n+1,\,2k}^2. \end{aligned}$$

18. Let $X(t)$ be standard Brownian motion, and for $\varepsilon > 0$ and $T > 1$ let $g_{\varepsilon,\,T}(x)$ be the conditional probability density for $X(1)$, given $X(t) \geq -\varepsilon$ for all $t \leq T$. Show

$$\lim_{\substack{\varepsilon \downarrow 0 \\ T \to \infty}} g_{\varepsilon,\,T}(x) = \sqrt{\frac{2}{\pi}}\, x^2 \exp(-x^2/2).$$

Remark: This is the distribution of $R(1)$ in a 3-dimensional Bessel process.

19. $\{f_\theta(X(t), t)\}$ is a martingale for any real parameter θ, where $f_\theta(x, t) = \exp\{\theta x - \frac{1}{2}\theta^2 t\}$. Use the martingale $f_\theta(X(t), t) + f_{-\theta}(X(t), t)$, where $\theta = \sqrt{2\lambda}$ to show

$$E[e^{-\lambda T}] = \frac{1}{\cosh(\sqrt{2\lambda}\, a)},$$

where $T = \min\{t\colon X(t) = +a \text{ or } X(t) = -a\}$.

20. Set

$$p(x, t) = \frac{1}{\sqrt{t}} \exp(x^2/2t), \qquad t > 0.$$

Show that $p(X(t), a + t)$ is a martingale for $a > 0$.

Hint: Verify

$$p(x, t) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} f_\theta(x, t)\, d\theta,$$

where $f_\theta(x, t) = \exp\{\theta x - \frac{1}{2}\theta^2 t\}$. Consult also (5.2).

21. Use the martingale in Problem 20 and the maximal inequality for martingales to show

$$\Pr\{|X(t)| \geq \sqrt{2(a+t)\log\sqrt{a+t}}, \quad \text{for some} \quad t \geq 0\} \leq \frac{1}{\sqrt{a}}.$$

22. Fix $a < 0 < b$ and let $T(a)$ [respectively, $T(b)$] be the first time the process reaches a (respectively, b). Let $I_a = 1$ if $T(a) < T(b)$, and zero otherwise. Similarly, let I_b be the indicator of the event that b is reached before a. Show that

$$\exp(-\sqrt{2\lambda}\, b) = E\{I_b \exp[-\lambda T(b)]\} + E\{I_a \exp[-\lambda T(a)]\} \exp[-\sqrt{2\lambda}\,(b-a)],$$

and

$$\exp(\sqrt{2\lambda}\, a) = E\{I_a \exp[-\lambda T(a)]\} + E\{I_b \exp[-\lambda T(b)]\} \exp[-\sqrt{2\lambda}\,(b-a)].$$

Hint: The first equation dichotomizes paths to b according to whether a is hit first or not. If a is hit first, a move to a without hitting b is followed by a move from a to b. Finally, recall

$$\exp(-\sqrt{2\lambda}\, b) = E[\exp[-\lambda T(b)] | X(0) = 0],$$

and

$$\exp[-\sqrt{2\lambda}\,(b-a)] = E[\exp[-\lambda T(b)] | X(0) = a].$$

23. (Continuation). Solve the equations derived in Problem 22 simultaneously for $E\{I_a \exp[-\lambda T(a)]\}$ and $E\{I_b \exp[-\lambda T(b)]\}$. Let $T = \min\{T(a), T(b)\}$ be the first time either a or b is reached. Find

$$E[\exp(-\lambda T)] = E\{I_a \exp[-\lambda T(a)]\} + E\{I_b \exp[-\lambda T(b)]\}.$$

24. Let $W(t)$ be a Brownian motion with positive drift $\mu > 0$ and variance σ^2. Let $M(t) = \max_{0 \leq u \leq t} W(u)$ and $Y(t) = M(t) - W(t)$. Fix $a > 0$ and $y > 0$, and let

$$T(a) = \min\{t\colon M(t) = a\}, \qquad S(y) = \min\{t\colon Y(t) = y\}.$$

Establish that

$$\Pr\{T(a) < S(y)\} = \exp\left\{\frac{-2\mu a}{\sigma^2[\exp(2\mu y/\sigma^2) - 1]}\right\}.$$

Hint: Let $f(a) = \Pr\{T(a) < S(y)\}$. Argue first, that $f(a_1 + a_2) = f(a_1)f(a_2)$, and thus $f(a) = e^{-ka}$ for some constant k.

With $\lambda = -2\mu/\sigma^2$, $e^{\lambda X(t)} = e^{\lambda[M(t) - Y(t)]}$ is a martingale. Set

$$T = \min\{T(a), S(y)\}$$

and apply the optional stopping theorem. Observe $M(T) = a$, $Y(T) = 0$, if $T(a) < S(y)$, and $Y(T) = y$ if $T(a) > S(y)$. Disect $1 = E[e^{\lambda[M(T) - Y(T)]}]$ according as $T(a) < S(y)$ or $T(a) > S(y)$. Let $a \to 0$ to determine the unknown constant k.

25. Let $\{X(t); t \geq 0\}$ be a Brownian motion process. By formally differentiating the martingale

$$\mathscr{Z}_\theta(t) = \exp\{\theta X(t) - (1/2)\theta^2 t\},$$

with respect to θ, show that, for each n, $H_n(X(t), t)$ is a martingale, where

$$H_0(x, t) \equiv 1,$$
$$H_1(x, t) = x,$$

and

$$H_n(x, t) = xH_{n-1}(x, t) - (n-1)tH_{n-2}(x, t).$$

An alternative approach is to show that (5.2) applies.

26. Consider any continuous integrable function f defined on the real line satisfying

$$\int_{-\infty}^{\infty} f(\delta)\, d\delta = a > 0.$$

Form the process

$$Y(t) = \frac{1}{\sqrt{t}} \int_0^t f(X(u))\, du.$$

Show that

$$\lim_{t\to\infty} E[Y(t)] \qquad \text{and} \qquad \lim_{t\to\infty} E[Y^2(t)]$$

exist and determine their values.

27. (Continuation.)
Show that

$$\lim_{t\to\infty} E[\{Y(t)\}^k] = \mu_k a^k$$

where μ_k is the kth moment of the one-sided normal distribution. (The one sided normal is the distribution of $|Z|$ where Z follows a standard normal distribution.)

NOTES

For applications of Brownian motion to statistical mechanics and mathematical analysis we recommend the delightful monograph by Kac [2].

An outstanding treatise on diffusion processes, which completes and profoundly extends the work of Lévy, is that of Ito and McKean [3].

REFERENCES

1. P. Lévy, "Processus Stochastiques et Mouvement Brownien." Gauthier-Villars, Paris, 1948.
2. M. Kac, "Probability and Related Topics in Physical Sciences." Wiley, New York, 1959.
3. K. Ito and H. P. McKean, "Diffusion Processes and Their Sample Paths." Springer-Verlag, Berlin, 1965.

Chapter 8

BRANCHING PROCESSES

The first four sections of this chapter provide a basic introduction to branching processes and their applications. Sections 5 through 11 provide generalizations and extensions and should not be attempted until after the earlier material has been mastered. In a one semester course, where time is scarce, these later sections might be omitted.

1: Discrete Time Branching Processes

Branching processes were introduced as examples of Markov chains in Section 2 of Chapter 2. There are numerous examples of Markov branching processes that arise naturally in various scientific disciplines. We list some of the more prominent ones.

(*a*) *Electron Multipliers*
An electron multiplier is a device that amplifies a weak current of electrons. A series of plates are set up in the path of electrons emitted by a source. Each electron, as it strikes the first plate, generates a random number of new electrons, which in turn strike the next plate and produce more electrons, etc. Let X_0 be the number of electrons initially emitted, X_1 the number of electrons produced on the first plate by the impact due to the X_0 initial electrons; in general let X_n be the number of electrons emitted from the nth plate due to electrons emanating from the $(n-1)$st plate. The sequence of random variables $X_0, X_1, X_2, \ldots, X_n, \ldots$ constitutes a branching process.

(*b*) *Neutron Chain Reaction*
A nucleus is split by a chance collision with a neutron. The resulting fission yields a random number of new neutrons. Each of these secondary

neutrons may hit some other nucleus producing a random number of additional neutrons, etc. In this case the initial number of neutrons is $X_0 = 1$. The first generation of neutrons comprises all those produced from the fission caused by the initial neutron. The size of the first generation is a random variable X_1. In general the population X_n at the nth generation is produced by the chance hits of the X_{n-1} individual neutrons of the $(n-1)$st generation.

(*c*) *Survival of Family Names*

The family name is inherited by sons only. Suppose that each individual has probability p_k of having k male offspring. Then from one individual there result the 1st, 2nd, ..., nth, ... generations of descendants. We may investigate the distribution of such random variables as the number of descendants in the nth generation, or the probability that the family name will eventually become extinct. Such questions will be dealt with in the general analysis of branching processes of this chapter.

(*d*) *Survival of Mutant Genes*

Each individual gene has a chance to give birth to k offspring, $k = 1, 2, \ldots$, which are genes of the same kind. However, any individual has a chance to transform into a different type or mutant gene. This gene may become the first in a sequence of generations of a particular mutant gene. We may inquire about the chances of survival of the mutant gene within the population of the original genes.

All of the above examples possess the following structure. Let X_0 denote the size of the initial population. Each individual gives birth, *independently of the others*, with probability p_k to k new individuals, where

$$p_k \geq 0, \qquad k = 0, 1, 2, \ldots, \qquad \text{and} \qquad \sum_{k=0}^{\infty} p_k = 1. \tag{1.1}$$

The totality of all the direct descendants of the initial population constitutes the first generation whose size we denote by X_1. Each individual of the first generation independently bears a progeny whose size is governed by the probability distribution (1.1). The descendants produced constitute the second generation of size X_2. In general the nth generation is composed of descendants of the $(n-1)$st generation each of whose members independently produces k progeny with probability p_k, $k = 0, 1, 2, \ldots$. The population size of the nth generation is denoted by X_n. The X_n form a sequence of integer-valued random variables which generate a Markov chain.

2: *Generating Function Relations for Branching Processes*

We will develop some relations for the probability generating functions of the X_n. Assume first that the initial population consists of one individual, i.e., assume $X_0 = 1$. Clearly we can write for every $n = 0, 1, 2, \ldots$

$$X_{n+1} = \sum_{r=1}^{X_n} \xi_r,$$

where ξ_r $(r \geq 1)$ are independently identically distributed random variables with distribution

$$\Pr\{\xi_r = k\} = p_k, \qquad k = 0, 1, 2, \ldots, \qquad \sum_{k=0}^{\infty} p_k = 1.$$

We introduce the probability generating function

$$\varphi(s) = \sum_{k=0}^{\infty} p_k s^k$$

and

$$\varphi_n(s) = \sum_{k=0}^{\infty} \Pr\{X_n = k\} s^k, \qquad \text{for} \quad n = 0, 1, 2, \ldots.$$

Manifestly,

$$\varphi_0(s) \equiv s \qquad \text{and} \qquad \varphi_1(s) = \varphi(s).$$

Further,

$$\begin{aligned}
\varphi_{n+1}(s) &= \sum_{k=0}^{\infty} \Pr\{X_{n+1} = k\} s^k \\
&= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \Pr\{X_{n+1} = k | X_n = j\} \Pr\{X_n = j\} s^k \\
&= \sum_{k=0}^{\infty} s^k \sum_{j=0}^{\infty} \Pr\{X_n = j\} \cdot \Pr\{\xi_1 + \cdots + \xi_j = k\} \\
&= \sum_{j=0}^{\infty} \Pr\{X_n = j\} \cdot \sum_{k=0}^{\infty} \Pr\{\xi_1 + \cdots + \xi_j = k\} s^k. \qquad (2.1)
\end{aligned}$$

Since ξ_r $(r = 1, 2, \ldots, j)$ are independent, identically distributed random variables with common probability generating function $\varphi(s)$, the sum $\xi_1 + \cdots + \xi_j$ has the probability generating function $[\varphi(s)]^j$. Thus,

$$\varphi_{n+1}(s) = \sum_{j=0}^{\infty} \Pr\{X_n = j\} [\varphi(s)]^j.$$

But the right-hand side is just the generating function $\varphi_n(\cdot)$ evaluated at $\varphi(s)$. Thus,

$$\varphi_{n+1}(s) = \varphi_n(\varphi(s)). \tag{2.2}$$

Iterating this relation we obtain

$$\begin{aligned}\varphi_{n+1}(s) &= \varphi_n(\varphi(s)) = \varphi_{n-1}(\varphi(\varphi(s))) = \varphi_{n-1}(\varphi_2(s)) \\ &= \varphi_{n-2}(\varphi_2(\varphi(s))) = \varphi_{n-2}(\varphi_3(s)).\end{aligned}$$

It follows, by induction, that for any $k = 0, 1, \ldots, n$

$$\varphi_{n+1}(s) = \varphi_{n-k}(\varphi_{k+1}(s)).$$

In particular, with $k = n - 1$,

$$\varphi_{n+1}(s) = \varphi(\varphi_n(s)). \tag{2.3}$$

If instead of $X_0 = 1$ we assume $X_0 = i_0$ (constant), then

$$\varphi_0(s) \equiv s^{i_0} \qquad \text{and} \qquad \varphi_1(s) = [\varphi(s)]^{i_0}$$

because

$$X_1 = \sum_{j=1}^{i_0} \xi_j.$$

We still have

$$\varphi_{n+1}(s) = \varphi_n(\varphi(s))$$

but (2.3) no longer holds.

With the help of (2.2), we will now compute the expectation and variance of X_n. It is assumed henceforth, unless explicitly stated to the contrary, that $X_0 = 1$. We postulate that

$$m = EX_1 \qquad \text{and} \qquad \sigma^2 = \operatorname{Var} X_1 = E(X_1^2) - [E(X_1)]^2$$

exist and are finite.

Obviously, $EX_n = \varphi_n'(1)$. Then differentiating (2.2) and setting $s = 1$ yields [since $\varphi(1) = 1$] $\varphi_{n+1}'(1) = \varphi_n'(1)\varphi'(1)$. Iteration produces

$$\varphi_{n+1}'(1) = \varphi'(1)\varphi_n'(1) = [\varphi'(1)]^2\varphi_{n-1}'(1) = [\varphi'(1)]^3\varphi_{n-2}'(1)$$

and by induction

$$\varphi_{n+1}'(1) = [\varphi'(1)]^n\varphi_1'(1) = [\varphi'(1)]^{n+1}.$$

But $\varphi'(1) = \varphi_1'(1) = EX_1 = m$. Thus

$$EX_{n+1} = m^{n+1}. \tag{2.4}$$

To compute $\operatorname{Var} X_{n+1}$, first note that

$$\varphi_n''(1) = \sum_{k=2}^{\infty} k(k-1)\Pr\{X_n = k\} = EX_n^2 - EX_n = EX_n^2 - \varphi_n'(1)$$

and so

$$\text{Var } X_n = \varphi_n''(1) + \varphi_n'(1) - [\varphi_n'(1)]^2.$$

But differentiating (2.3) twice and setting $s = 1$ yields

$$\varphi_{n+1}''(1) = \varphi''(1)[\varphi_n'(1)]^2 + \varphi'(1)\varphi_n''(1).$$

Since $\varphi'(1) = m$ and $\varphi''(1) = EX_1^2 - EX_1 = \sigma^2 + m^2 - m$, we have

$$\varphi_{n+1}''(1) = Mm^{2n} + m\varphi_n''(1),$$

where $M = \sigma^2 + m^2 - m$. By induction,

$$\varphi_{n+1}''(1) = M\{m^{2n} + m^{2n-1}\} + m^2\varphi_{n-1}''(1) = \cdots = M\{m^{2n} + m^{2n-1} + \cdots + m^n\}.$$

Thus

$$\begin{aligned}\text{Var } X_{n+1} &= (\sigma^2 + m^2 - m)\{m^{2n} + m^{2n-1} + \cdots + m^n\} + m^{n+1} - m^{2n+2} \\ &= \sigma^2\{m^{2n} + m^{2n-1} + \cdots + m^n\} \\ &= \sigma^2 m^n \frac{m^{n+1} - 1}{m - 1} \qquad \text{if} \quad m \neq 1\end{aligned}$$

and

$$\text{Var } X_{n+1} = (n+1)\sigma^2 \qquad \text{if} \quad m = 1.$$

We have hereby verified the formulas $EX_n = m^n$ and

$$\text{Var } X_n = \begin{cases} \sigma^2 m^{n-1} \dfrac{m^n - 1}{m - 1} & \text{if} \quad m \neq 1, \\ n\sigma^2 & \text{if} \quad m = 1. \end{cases}$$

Thus, the variance increases (decreases) geometrically if $m > 1$ ($m < 1$), and linearly if $m = 1$. This behavior is characteristic of many results for branching processes.

3: Extinction Probabilities

We want to determine the probability that the population will eventually die out, i.e., $\Pr\{X_n = 0 \text{ for some } n\}$. Obviously whenever $X_n = 0$, $X_k = 0$ for all $k > n$.

Note first that extinction never occurs if the probability that an individual gives birth to no offspring is zero, i.e., when $p_0 = 0$. Thus in investigating the probability of extinction we will assume $0 < p_0 < 1$. Let

$$q_n = \Pr\{X_n = 0\} = \varphi_n(0).$$

Then by formula (2.3)

$$q_{n+1} = \varphi_{n+1}(0) = \varphi(\varphi_n(0)) = \varphi(q_n). \tag{3.1}$$

Since $\varphi(s)$ is a strictly increasing function (it is a power series with non-negative coefficients and $p_0 < 1$) and $q_1 = \varphi_1(0) = p_0 > 0$, $q_2 = \varphi(q_1) > \varphi(0) = q_1$. Assume that $q_n > q_{n-1}$; then $q_{n+1} = \varphi(q_n) > \varphi(q_{n-1}) = q_n$. This shows inductively that $q_1, q_2, \ldots, q_n, \ldots$ is a monotone increasing sequence bounded by 1. Hence

$$\pi = \lim_{n\to\infty} q_n$$

exists and $0 < \pi \leq 1$. Since $\varphi(s)$ is continuous, for $0 \leq s \leq 1$ [at $s = 1$, by Abel's lemma (Lemma 5.1, Chapter 2)], letting $n \to \infty$ in (3.1) yields

$$\pi = \varphi(\pi). \tag{3.2}$$

Since q_n is defined as the probability of extinction at or prior to the nth generation, we infer that π is the probability of eventual extinction and (3.2) shows that π is a root of the equation

$$\varphi(s) = s. \tag{3.3}$$

We now establish the result that π is the smallest positive root of (3.3). Let s_0 be a positive root of (3.3). Then $q_1 = \varphi(0) < \varphi(s_0) = s_0$. Assume $q_n < s_0$. Then by (3.1) $q_{n+1} = \varphi(q_n) < \varphi(s_0) = s_0$. Thus we infer by induction that $q_n < s_0$ holds for all n. It follows that $\pi = \lim q_n \leq s_0$, validating the assertion that π is the smallest positive root of (3.3).

Now, assume also that $p_0 + p_1 < 1$. Then $\varphi(s)$ is a convex function in $0 < s \leq 1$, as $\varphi''(s) = \sum_{k=2}^{\infty} k(k-1)p_k s^{k-2} > 0$. Therefore, the graph of $\varphi(s)$ can intersect the 45° line in at most two points. We know that $\varphi(1) = 1$ and so there certainly is an intersection at (1, 1). Clearly we have one of the two cases represented by Figs. 1 and 2. If $m = \varphi'(1) > 1$, then the slope of the tangent to the graph of $\varphi(s)$ at $s = 1$ exceeds 1 and the case represented by Fig. 1 is germane. In this case $0 < \pi < 1$. If $m = \varphi'(1) \leq 1$, then the slope of the tangent at $s = 1$ is smaller than or equal to one and we are in the situation of Fig. 2. Then necessarily $\pi = 1$. Thus, we have proved that the probability of extinction is 1 if $m \leq 1$ and is less than 1 if $m > 1$. In other words, extinction is certain if and only if the mean number of offspring per individual does not exceed one.

Further, note that for $0 \leq s \leq \pi$, $\varphi(s) \leq \varphi(\pi)$ (see Fig. 2). By induction we have $\varphi_n(s) \leq \pi$ $(0 \leq s \leq \pi)$ for all n. But $\varphi_n(s) \geq \varphi_n(0) = q_n$ and thus, $q_n \leq \varphi_n(s) \leq \pi$. Let $n \to \infty$. Then

$$\lim_{n\to\infty} \varphi_n(s) = \pi \qquad \text{for} \quad 0 \leq s \leq \pi.$$

Further, for the case $m > 1$, when $\pi < s < 1$ we have $\pi < \varphi(s) < s < 1$ (consult Fig. 1). By induction

$$\pi < \varphi_n(s) < \varphi_{n-1}(s) < \cdots \qquad (\pi < s < 1).$$

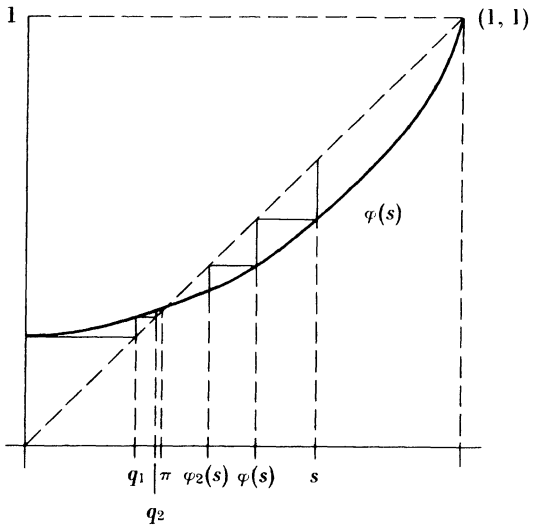


FIG. 1

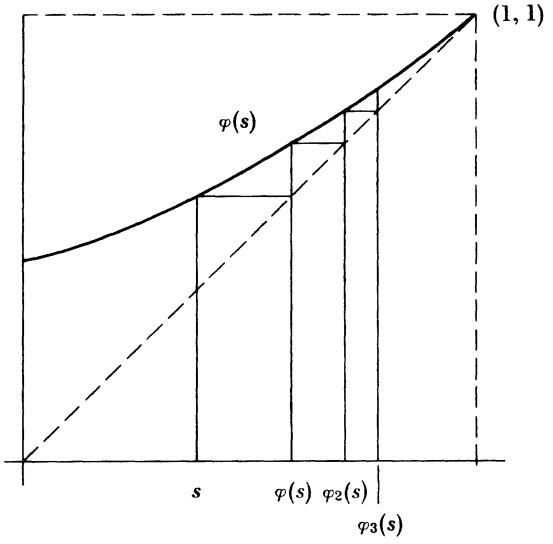


FIG. 2

It follows that

$$\lim_{n\to\infty} \varphi_n(s) \geq \pi. \tag{3.4}$$

The limit has to equal π for if $\lim_{n\to\infty} \varphi_n(s) = \alpha > \pi$, then $\varphi(\alpha) < \alpha$ and the convergence in (3.4) is impossible in view of the relation $\lim_{n\to\infty} \varphi_{n+1}(s) = \lim_{n\to\infty} \varphi(\varphi_n(s))$. The above analysis shows that

$$\lim_{n\to\infty} \varphi_n(s) = \pi \qquad \text{for} \quad 0 \leq s < 1.$$

The fact that $\varphi_n(s)$ converge to the constant function π on $0 \leq s < 1$ implies that in the series

$$\varphi_n(s) = \sum_{k=0}^{\infty} \Pr\{X_n = k\} s^k$$

the first coefficient

$$\Pr\{X_n = 0\} \quad \text{converges to} \quad \pi \qquad \text{as} \quad n \to \infty,$$

and all the other coefficients

$$\Pr\{X_n = k\} \quad \text{converge to} \quad 0 \qquad \text{as} \quad n \to \infty \quad \text{for} \quad k = 1, 2, \ldots.$$

Hence, regardless of the actual value of $m = EX_1 > 1$, the probability that the nth generation will consist of any positive finite number of individuals tends to zero as $n \to \infty$, while the probability of extinction tends to π. In this circumstance we say that $X_n \to \infty$ as $n \to \infty$ with probability $1 - \pi$.

This result is also a consequence of the general theory of Markov chains in that the Markov chain determined by the sequence $X_0, X_1, X_2, \ldots$ has a single absorbing state $\{0\}$ and so $\lim_{n\to\infty} P_{ij}^n = 0$, $1 \leq i, j < \infty$, since i and j are automatically transient.

We close this section by noting the interesting property that the conditional expectation of X_{n+r} (r a positive integer), given X_n, is $m^r \cdot X_n$, i.e., $E(X_{n+r} | X_n) = m^r X_n$. To prove this we first consider the case $r = 1$:

$$E\left\{X_{n+1} \,\middle|\, X_n\right\} = E\left\{\sum_{j=1}^{X_n} \xi_j \,\middle|\, X_n\right\} = X_n E\xi_j = mX_n.$$

We now assume the stated relation for r and prove the formula for $r + 1$. Thus

$$E\{X_{n+r+1} | X_n\}$$
$$= E\{E[X_{n+r+1} | X_{n+r}, X_{n+r-1}, \ldots, X_n] | X_n\} = E\{E[X_{n+r+1} | X_{n+r}] | X_n\},$$

where we use the Markov nature of $\{X_n\}$. But $E[X_{n+r+1} | X_{n+r}] = X_{n+r} \cdot m$ and by the induction hypothesis, $E(mX_{n+r} | X_n) = m^{r+1} X_n$. Thus we have

$$E\{X_{n+r} | X_n\} = X_n m^r \qquad \text{for} \quad r = 0, 1, 2, \ldots, \quad n = 0, 1, 2, \ldots. \tag{3.5}$$

Now consider the random variables

$$W_n = \frac{X_n}{m^n} \qquad n = 0, 1, 2, \ldots.$$

Then on the basis of (3.5) , we have

$$E\{W_{n+r}|W_n\} = \frac{1}{m^{n+r}} E\{X_{n+r}|X_n\} = \frac{1}{m^{n+r}} \cdot X_n \cdot m^r = W_n.$$

We may write for $r, n = 0, 1, 2, \ldots$

$$E\{W_{n+r}|W_n, W_{n-1}, \ldots, W_1, W_0\} = W_n, \tag{3.6}$$

which shows that $\{W_n\}_{n=0}^{\infty}$ is a martingale.

4: Examples

(i) Let $\varphi(s) = p_0 + p_1 s$, $0 < p_0 < 1$. The associated branching process is a pure death process. In each period each individual independently dies with probability p_0 and survives with probability $1 - p_0 = p_1$.

(ii) Let $\varphi(s) = p_0 + p_2 s^2$ $(0 < p_0 < 1,\ p_0 + p_2 = 1)$. This is the probability-generating function corresponding to a branching process in which in each generation an individual either dies or is replaced by two progeny.

(iii) Consider the example where each individual produces N or 0 direct descendants with probabilities p or q respectively. Thus $p_0 = q$, $p_N = p$, and $p_k = 0$ for $k \neq 0, N$. Then

$$\varphi(s) = q + ps^N. \tag{4.1}$$

(iv) Each individual may have k offspring where k has a binomial probability distribution with parameters N and p. Then

$$\varphi(s) = (q + ps)^N. \tag{4.2}$$

(v) In connection with Example (d) described at the beginning of this chapter it is frequently assumed that the probability of a mutant gene having k direct descendants $(k = 0, 1, 2, \ldots)$ is governed by a Poisson distribution with mean $\lambda = 1$. Then $\varphi(s) = e^{s-1}$ and $\pi = 1$.

The rationale behind this choice of distribution is as follows. In many populations a large number of zygotes (fertilized eggs) are produced, only a small number of which grow to maturity. The events of fertilization and maturation of different zygotes obey the law of independent binomial trials. The number of trials (i.e., number of zygotes) is large so that the actual number of mature progeny follows the Poisson distribution. This is a corollary of the principle of rare events commonly invoked to justify the Poisson approximation. It seems quite appropriate in the model of population growth of a rare mutant gene. If the mutant gene carries a

biological advantage (or disadvantage), then the probability distribution is taken to be the Poisson distribution with mean $\lambda > 1$ or (< 1). Specifically,

$$\varphi(s) = e^{\lambda(s-1)} \tag{4.3}$$

and $0 < \pi < 1$ if and only if $\lambda > 1$.

In a heterogeneous population of mutant genes we may assume that the probability distribution of the number of offspring is a Poisson distribution, but with the mean also a random variable.

For example, we may have a large geographical area in which for each subarea a branching process characterized by the probability generating function of a Poisson distribution with parameter λ is taking place. We assume, furthermore, that the value of λ varies depending on the subarea and its distribution over the whole area is that of a gamma. Formally we postulate that the probability of a mutant gene having exactly k direct descendants is given by

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots,$$

where λ itself is a random variable distributed according to a gamma distribution with the density function

$$f(\lambda) = \frac{(q/p)^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} \exp\left(-\frac{q}{p}\lambda\right) \qquad \text{for} \quad \lambda \geq 0,$$

$$= 0 \qquad \text{otherwise,}$$

where q, p, α are positive constants and $q + p = 1$. The probability of an individual having k offspring, if we average with respect to the values of the parameter λ, is

$$\Pr\{\xi = k\} = \int_0^\infty \Pr\{\xi = k|\lambda\} f(\lambda)\, d\lambda.$$

Thus the generating function is

$$\varphi(s) = \sum_{k=0}^\infty \Pr\{\xi = k\} s^k = \sum_{k=0}^\infty \int_0^\infty \exp(-\lambda)\frac{\lambda^k}{k!}\frac{(q/p)^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} \exp\left(-\frac{q}{p}\lambda\right) d\lambda \cdot s^k$$

$$= \int_0^\infty \exp(-\lambda)\frac{(q/p)^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} \exp\left(-\frac{q}{p}\lambda\right)\left(\sum_{k=0}^\infty \frac{(\lambda s)^k}{k!}\right) d\lambda$$

$$= \int_0^\infty \exp\left\{-\left(\frac{q}{p} + 1 - s\right)\lambda\right\} \frac{(q/p)^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)}\, d\lambda$$

$$= \left(\frac{q/p}{(q/p) + 1 - s}\right)^\alpha = \left(\frac{q}{1 - ps}\right)^\alpha.$$

This we recognize as the probability generating function of the negative binomial distribution.

(vi) In Examples (ii)–(iv) no closed-form expressions are known for the nth-generation probability generating function $\varphi_n(s)$. The final example studied below is amenable to a rather complete analysis. Specifically, we will compute the nth-generation probability generating function. Let

$$p_k = bc^{k-1}, \qquad k = 1, 2, \ldots,$$

and

$$p_0 = 1 - \sum_{k=1}^{\infty} p_k,$$

where $b, c > 0$ and $b + c < 1$. Then

$$p_0 = 1 - b \sum_{k=1}^{\infty} c^{k-1} = 1 - \frac{b}{1-c} = \frac{1-b-c}{1-c}$$

and the corresponding probability generating function is

$$\varphi(s) = 1 - \frac{b}{1-c} + bs \sum_{k=1}^{\infty} (cs)^{k-1} = \frac{1-(b+c)}{1-c} + \frac{bs}{1-cs}. \tag{4.4}$$

Notice that $\varphi(s)$ has the form of a linear fractional transformation

$$f(s) = \frac{\alpha + \beta s}{\gamma + \delta s}, \qquad \alpha\delta - \beta\gamma \neq 0. \tag{4.5}$$

We now record several elementary properties of linear fractional transformations needed below:

(i) Iterates of linear fractional transformations are again linear fractional transformations, for if $f(s)$ is defined by (4.5) simple algebra gives

$$f(f(s)) = \frac{\alpha(\gamma + \beta) + (\alpha\delta + \beta^2)s}{\alpha\delta + \gamma^2 + \delta(\gamma + \beta)s}.$$

(ii) There always exist two finite (possibly identical) solutions to the equation $f(s) = s$. The solutions are called fixed points of $f(\cdot)$. If $f(s)$ is a probability-generating function then $s_1 = 1$ is one of the fixed points and we shall see that the other fixed point s_0 is less than one, equal to one, or greater than one, according to whether $f'(1)$ is greater than, equal to, or less than one.

For the generating function given by (4.4), one can verify by straightforward algebra that the second fixed point, for $c > 0$, and $b + c < 1$, is

$$s_0 = \frac{1-b-c}{c(1-c)}.$$

(iii) For any two points s_i, $i = 0, 1$, it is easily seen that

$$\frac{f(s)-f(s_i)}{s-s_i} = \frac{\gamma\beta - \alpha\delta}{(\gamma+\delta s)(\gamma+\delta s_i)}.$$

Hence

$$\frac{f(s)-f(s_0)}{f(s)-f(s_1)} = \left(\frac{\gamma+\delta s_1}{\gamma+\delta s_0}\right)\left(\frac{s-s_0}{s-s_1}\right). \tag{4.6}$$

If we now let s_0 and s_1 be the two (nonidentical) fixed points of $f(\cdot)$ and write $w = f(s)$, (4.6) becomes

$$\frac{w-s_0}{w-s_1} = \kappa\frac{s-s_0}{s-s_1}, \tag{4.7}$$

where κ can be calculated from (4.6) or more simply from (4.5) by setting $s = 0$.

Using (4.7) we easily obtain the iterates $f_n(s) = w_n$ of $f(s)$:

$$\frac{w_2-s_0}{w_2-s_1} = \kappa\frac{w_1-s_0}{w_1-s_1} = \kappa\left(\kappa\frac{s-s_0}{s-s_1}\right),$$

and in general

$$\frac{w_n-s_0}{w_n-s_1} = \kappa^n\frac{s-s_0}{s-s_1}. \tag{4.8}$$

For the generating function of the geometric distribution given by (4.4), noting that the fixed points are $s_0 = (1-b-c)/c(1-c)$ and $s_1 = 1$, we obtain

$$\kappa = \frac{(1-c)^2}{b} = \frac{1}{m},$$

where m is the mean of the geometric distribution. For $m \neq 1$ the two fixed points s_0 and 1 are different; hence, solving for w_n in (4.8) gives

$$w_n = \frac{s_0 - (1/m^n)[(s-s_0)/(s-1)]}{1-(1/m^n)[(s-s_0)/(s-1)]}, \qquad m \neq 1, \tag{4.9}$$

which may be written in the form

$$\varphi_n(s) = 1 - m^n\left(\frac{1-s_0}{m^n-s_0}\right) + \frac{m^n[(1-s_0)/(m^n-s_0)]^2 s}{1-[(m^n-1)/(m^n-s_0)]s} \tag{4.10}$$

Then the probabilities of extinction at the nth generation are

$$\Pr\{X_n = 0\} = \varphi_n(0) = 1 - m^n\left(\frac{1-s_0}{m^n-s_0}\right).$$

Note that this expression converges to s_0 as $n \to \infty$ if $m > 1$ and to 1 if $m < 1$. The probability of a given population size in the nth generation, $\Pr\{X_n = k\}$, $k = 1, 2, \ldots$, can be computed by simply expanding (4.10) as a power series in s. If we define the time to extinction T as the smallest subscript n such that $X_n = 0$, i.e., the first passage time into state 0, then

$$\Pr\{T \leq n\} = \Pr\{X_n = 0\} = \varphi_n(0)$$

and

$$\Pr\{T = n\} = \Pr\{T \leq n\} - \Pr\{T \leq n-1\} = \varphi_n(0) - \varphi_{n-1}(0).$$

In the case $m \neq 1$, we have

$$\Pr\{T = n\} = 1 - m^n \left(\frac{1-s_0}{m^n - s_0}\right) - 1 + m^{n-1}\left(\frac{1-s_0}{m^{n-1} - s_0}\right)$$

$$= m^{n-1} s_0 \frac{(m-1)(1-s_0)}{(m^n - s_0)(m^{n-1} - s_0)} \qquad \text{for} \quad n = 1, 2, \ldots.$$

If $m = 1$, then $b = (1-c)^2$ and the equation $\varphi(s) = s$ has the double root $s = 1$ and no other root. In fact,

$$\varphi(s) = c + \frac{(1-c)^2 s}{1 - cs} = \frac{c - (2c-1)s}{1 - cs}.$$

Then

$$\varphi_2(s) = \varphi(\varphi(s)) = \frac{c - (2c-1)[(c-(2c-1)s)/(1-cs)]}{1 - c[(c-(2c-1)s)/(1-cs)]} = \frac{2c - (3c-1)s}{1 + c - 2cs}$$

and by induction

$$\varphi_n(s) = \frac{nc - [(n+1)c - 1]s}{1 + (n-1)c - ncs}. \tag{4.11}$$

In the case $m = 1$ we have the extinction probabilities

$$\Pr\{X_n = 0\} = \varphi_n(0) = \frac{nc}{1 + (n-1)c} \qquad \text{for} \quad n = 1, 2, \ldots.$$

Further, the time to extinction T has the distribution

$$\Pr\{T = n\} = \frac{nc}{1 + (n-1)c} - \frac{(n-1)c}{1 + (n-2)c} = \frac{c(1-c)}{[1 + (n-1)c]\,[1 + (n-2)c]}$$

5: Two-Type Branching Processes

We generalize the previous developments to two dimensions. Consider a population of organisms or objects where two different types may be distinguished. Individuals of either type will produce offspring of possibly

both types independently. Let U_n and V_n be the number of individuals of types I and II, respectively, in the nth generation. We may write

$$U_{n+1} = \sum_{j=1}^{U_n} \xi_j^{(1)} + \sum_{j=1}^{V_n} \xi_j^{(2)},$$

$$V_{n+1} = \sum_{j=1}^{U_n} \zeta_j^{(1)} + \sum_{j=1}^{V_n} \zeta_j^{(2)},$$

where $(\xi_j^{(i)}, \zeta_j^{(i)})$ are independent, identically distributed, random vectors with distribution

$$\Pr\{\xi_j^{(i)} = k, \zeta_j^{(i)} = l\} = p_i(k, l), \qquad k, l = 0, 1, 2, \ldots,$$
$$\text{for} \quad j = 1, 2, \ldots \quad \text{and} \quad i = 1, 2.$$

Here $p_i(k, l) \geq 0$ and $\sum_{k,l=0}^{\infty} p_i(k, l) = 1$ for $i = 1, 2$.

In other words $p_1(k, l)$ and $p_2(k, l)$ are the probabilities that a single individual of type I and type II, respectively, produces $k + l$ direct descendants of which k are of type I and l are of type II.

We assume the process begins with a single individual, i.e., we assume either

$$U_0 = 1 \quad \text{and} \quad V_0 = 0 \tag{5.1}$$

or

$$U_0 = 0 \quad \text{and} \quad V_0 = 1. \tag{5.2}$$

We introduce the pair of two-dimensional probability generating functions

$$\varphi^{(i)}(s, t) = \sum_{k,l=0}^{\infty} p_i(k, l) s^k \cdot t^l, \qquad i = 1, 2,$$

that is,

$$\varphi_n^{(1)}(s, t) = \sum_{k,l=0}^{\infty} \Pr\{U_n = k, V_n = l | U_0 = 1, V_0 = 0\} s^k \cdot t^l,$$

$$\varphi_n^{(2)}(s, t) = \sum_{k,l=0}^{\infty} \Pr\{U_n = k, V_n = l | U_0 = 0, V_0 = 1\} s^k \cdot t^l.$$

The generating function of (5.1) is

$$\varphi_0^{(1)}(s, t) \equiv s,$$

and that of (5.2) is

$$\varphi_0^{(2)}(s, t) \equiv t.$$

Also

$$\varphi_1^{(i)}(s, t) = \varphi^{(i)}(s, t) \qquad \text{for} \quad i = 1, 2.$$

It can be shown by generalizing the method used for the one-dimensional process that

$$\varphi^{(i)}_{n+m}(s, t) = \varphi^{(i)}_m(\varphi^{(1)}_n(s, t), \varphi^{(2)}_n(s, t)), \quad \text{for } i = 1, 2 \text{ and } n, m = 0, 1, 2, \ldots. \tag{5.3}$$

This is the two-dimensional equivalent of formula (2.3).

To generalize formula (3.5) we introduce the following notation. Let $\mathbf{X}_n = (U_n, V_n)$ be the two-dimensional vector with components U_n and V_n. Let

$$m_{11} = E\{U_1 | U_0 = 1, V_0 = 0\} = E\xi^{(1)},$$
$$m_{12} = E\{V_1 | U_0 = 1, V_0 = 0\} = E\zeta^{(1)},$$
$$m_{21} = E\{U_1 | U_0 = 0, V_0 = 1\} = E\xi^{(2)}$$
$$m_{22} = E\{V_1 | U_0 = 0, V_0 = 1\} = E\zeta^{(2)}$$

and introduce the matrix of expectations

$$\mathbf{M} = \begin{Vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{Vmatrix}.$$

Thus m_{11} and m_{12} are the expected numbers of offspring of type I or type II, respectively, produced by a single parent of type I. Then as a generalization of (3.5) we have the matrix identity

$$E[\mathbf{X}_{n+r} | \mathbf{X}_n] = \mathbf{X}_n \mathbf{M}^r \quad \text{for } r, n = 0, 1, 2, \ldots. \tag{5.4}$$

The proof for $r = 1$ proceeds directly. Thus

$$E[\mathbf{X}_{n+1} | \mathbf{X}_n] =$$
$$\left(E\left[\sum_{j=1}^{U_n} \xi^{(1)}_j + \sum_{j=1}^{V_n} \xi^{(2)}_j | U_n, V_n) \right], \quad E\left[\sum_{j=1}^{U_n} \zeta^{(1)}_j + \sum_{j=1}^{V_n} \zeta^{(2)}_j | (U_n, V_n) \right] \right)$$
$$= (m_{11} U_n + m_{21} V_n, m_{12} U_n + m_{22} V_n)$$
$$= (U_n, V_n) \begin{Vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{Vmatrix}$$
$$= \mathbf{X}_n \cdot \mathbf{M}.$$

We now assume that relation (5.4) holds for r and prove it for $r + 1$. By the Markov property for $\{\mathbf{X}_n\}$, we have

$$\begin{aligned} E[\mathbf{X}_{n+r+1} | \mathbf{X}_n] &= E\{E[\mathbf{X}_{n+r+1} | \mathbf{X}_{n+r}, \ldots, \mathbf{X}_n] | \mathbf{X}_n\} \\ &= E\{E[\mathbf{X}_{n+r+1} | \mathbf{X}_{n+r}] | \mathbf{X}_n\} = E\{\mathbf{X}_{n+r} \mathbf{M} | \mathbf{X}_n\} \\ &= E\{\mathbf{X}_{n+r} | \mathbf{X}_n\} \cdot \mathbf{M} \quad \text{(using the induction hypothesis)} \\ &= \mathbf{X}_n \mathbf{M}^{r+1}. \end{aligned}$$

This completes the induction step.

For the two-dimensional branching process we may introduce the extinction probabilities

$$\pi^{(1)} = \Pr\{U_n = V_n = 0 \quad \text{for some} \quad n | U_0 = 1,\ V_0 = 0\},$$
$$\pi^{(2)} = \Pr\{U_n = V_n = 0 \quad \text{for some} \quad n | U_0 = 0,\ V_0 = 1\}.$$

The one-dimensional theory extends to this case with the remark that the role of the expectation m is played here by the largest eigenvalue ρ of the expectation matrix $\mathbf{M}$.

We direct the reader to the Appendix and particularly to the Frobenius theorem concerning matrices with nonnegative elements. It is proved there that if $\mathbf{M}$ is a matrix with positive elements (symbolically written here as $\mathbf{M} \gg 0$) then the eigenvalue of largest magnitude is positive and real. This eigenvalue is designated as $\rho(\mathbf{M}) = \rho$.

It is convenient to introduce the vector notations

$$\mathbf{u} = (s, t),$$
$$\boldsymbol{\phi}(\mathbf{u}) = (\varphi^{(1)}(s, t), \varphi^{(2)}(s, t)),$$
$$\boldsymbol{\phi}_n(\mathbf{u}) = (\varphi_n^{(1)}(s, t), \varphi_n^{(2)}(s, t)),$$
$$\boldsymbol{\pi} = (\pi^{(1)}, \pi^{(2)}),$$
$$\mathbf{1} = (1, 1).$$

Then we may state

Theorem 5.1. *Assume that the components of* $\boldsymbol{\phi}(\mathbf{u})$ *are not linear functions of* s *and* t *and that* $\mathbf{M} \gg \mathbf{0}$ *(every element of* $\mathbf{M}$ *is positive). Then* $\boldsymbol{\pi} = \mathbf{1}$ *if the largest eigenvalue* ρ *of* $\mathbf{M}$ *does not exceed one and* $\boldsymbol{\pi} \ll \mathbf{1}$ *if* $\rho > 1$. *(The notation* $\mathbf{u} \ll \mathbf{v}$ $(\mathbf{u} \leq \mathbf{v})$ *signifies that* $\mathbf{v} - \mathbf{u}$ *has positive (nonnegative) components.) In the case* $\rho > 1$, $\boldsymbol{\pi}$ *is the smallest nonnegative solution of*

$$\mathbf{u} = \boldsymbol{\phi}(\mathbf{u}), \qquad \mathbf{u} \ll \mathbf{1}. \tag{5.5}$$

Proof. Consider the case $\rho \leq 1$. According to the general theory of Markov chains we know that if a chain has a single absorbing state then all states from which the absorbing state may be reached are transient. The two-dimensional process $\mathbf{X}_n = (U_n, V_n)$ is just such a process: the origin is the only absorbing state and it may be reached from all other states. This is a consequence of the fact that $\boldsymbol{\phi}(\mathbf{u})$ has no linear components and $\rho \leq 1$. Thus every state with the exception of the origin is transient. Therefore, if $|\mathbf{X}_n| = U_n + V_n$, then

$$\Pr\{0 < |\mathbf{X}_n| < N \text{ for infinitely many } n\} = 0 \qquad \text{for any positive } N$$

(cf. Theorem 7.1 of Chapter 2). This means that

$$\Pr\{|\mathbf{X}_n| \to 0\} + \Pr\{|\mathbf{X}_n| \to \infty\} = 1.$$

From formula (5.4) $E[\mathbf{X}_n|\mathbf{X}_0] = \mathbf{X}_0\mathbf{M}^n$. But Theorem 2.3 of the Appendix asserts that $(1/\rho^n)\mathbf{M}^n$ converges componentwise as $n \to \infty$. Consequently, in the case $\rho \leq 1$, the components of $E[\mathbf{X}_n|\mathbf{X}_0]$ stay bounded as $n \to \infty$. It follows from this that the event $|\mathbf{X}_n| \to \infty$ occurs with probability zero. Hence $\Pr\{|\mathbf{X}_n| \to 0\} = 1$ or, what is the same, $U_n \to 0$ and $V_n \to 0$ as $n \to \infty$ with probability one. Thus if $\rho \leq 1$,

$$\pi^{(1)} = \pi^{(2)} = 1$$

holds.

Next consider the case $\rho > 1$. From formula (5.3) we have with $s = t = 0$

$$\varphi_{n+1}^{(i)}(0, 0) = \varphi^{(i)}(\varphi_n^{(1)}(0, 0), \varphi_n^{(2)}(0, 0)), \qquad i = 1, 2. \tag{5.6}$$

Let

$$q_n^{(1)} = \varphi_n^{(1)}(0, 0) = \Pr\{U_n = V_n = 0 | U_0 = 1, V_0 = 0\},$$
$$q_n^{(2)} = \varphi_n^{(2)}(0, 0) = \Pr\{U_n = V_n = 0 | U_0 = 0, V_0 = 1\}.$$

Then (5.6) is the same as

$$q_{n+1}^{(i)} = \varphi^{(i)}(q_n^{(1)}, q_n^{(2)}), \qquad i = 1, 2. \tag{5.7}$$

Since $\varphi^{(i)}(s, t)$ is increasing in the variables s and t, strictly so if both increase, and since $q_1^{(i)} = \varphi_1^{(i)}(0, 0) > 0$, $i = 1, 2$, we plainly have

$$q_2^{(i)} = \varphi^{(i)}(q_1^{(1)}, q_1^{(2)}) > \varphi^{(i)}(0, 0) = q_1^{(i)}, \qquad i = 1, 2.$$

Then by induction

$$q_{n+1}^{(i)} = \varphi^{(i)}(q_n^{(1)}, q_n^{(2)}) > \varphi^{(i)}(q_{n-1}^{(1)}, q_{n-1}^{(2)}) = q_n^{(i)}, \qquad i = 1, 2.$$

Hence, $q_n^{(i)}$, $n = 1, 2, 3, \ldots$, for each $i = 1, 2$, form a monotone increasing sequence bounded above by 1, and

$$\lim_{n \to \infty} q_n^{(i)} = \pi^{(i)} \leq 1, \qquad i = 1, 2.$$

Let $n \to \infty$ in (5.7). Then

$$\pi^{(i)} = \varphi^{(i)}(\pi^{(1)}, \pi^{(2)}), \qquad i = 1, 2,$$

or in vector notation $\boldsymbol{\pi} = \boldsymbol{\phi}(\boldsymbol{\pi})$. We will now prove that $\boldsymbol{\pi} \ll \mathbf{1}$ and that this is the unique solution of (5.5) under the conditions stated. Expanding $\varphi_n^{(i)}(\cdot, \cdot)$ according to Taylor's theorem about (1, 1) we have

$$\varphi_n^{(i)}(1 - s, 1 - t) = \varphi_n^{(i)}(1, 1) - \left(\frac{\partial \varphi_n^{(i)}(s, t)}{\partial s}\bigg|_{s=t=1}\right)s$$
$$- \left(\frac{\partial \varphi_n^{(i)}(s, t)}{\partial t}\bigg|_{s=t=1}\right)t + o(|s| + |t|), \tag{5.8}$$

which is valid for $|1-s| \leq 1$, $|1-t| \leq 1$, and s and t sufficiently small. The $o(\cdot)$ symbol signifies that $[o(|s|+|t|)]/(|s|+|t|) \to 0$ whenever $|s|+|t| \to 0$. Moreover,

$$\left.\frac{\partial \varphi_n^{(1)}(s,t)}{\partial s}\right|_{s=t=1} = E[U_n | U_0 = 1, V_0 = 0] = m_{11}^{(n)},$$

$$\left.\frac{\partial \varphi_n^{(1)}(s,t)}{\partial t}\right|_{s=t=1} = E[V_n | U_0 = 1, V_0 = 0] = m_{12}^{(n)},$$

$$\left.\frac{\partial \varphi_n^{(2)}(s,t)}{\partial s}\right|_{s=t=1} = E[U_n | U_0 = 0, V_0 = 1] = m_{21}^{(n)},$$

$$\left.\frac{\partial \varphi_n^{(2)}(s,t)}{\partial t}\right|_{s=t=1} = E[V_n | U_0 = 0, V_0 = 1] = m_{22}^{(n)}.$$

We may write (5.8) in vector form as

$$\boldsymbol{\varphi}_n(\mathbf{1}-\mathbf{u}) = \mathbf{1} - \mathbf{M}^{(n)}\mathbf{u} + \mathbf{o}(|s|+|t|), \tag{5.9}$$

where

$$\mathbf{M}^{(n)} = \begin{Vmatrix} m_{11}^{(n)} & m_{12}^{(n)} \\ m_{21}^{(n)} & m_{22}^{(n)} \end{Vmatrix}$$

and $|\mathbf{u}| < \varepsilon$. Of course $\mathbf{M}^{(n)} = \mathbf{M}^n$ as is evident from the relation $E[\mathbf{X}_n|\mathbf{X}_0] = \mathbf{X}_0\mathbf{M}^n$. Let the absolute value of a vector $\mathbf{v} = (v_1, v_2)$ be defined as the sum of the absolute values of its coordinates: $|\mathbf{v}| = |v_1| + |v_2|$. We will now prove that for n sufficiently large

$$|\mathbf{M}^n\mathbf{u}| > 2|\mathbf{u}|, \qquad \mathbf{u} = (s,t), \tag{5.10}$$

provided $\mathbf{u} \geq \mathbf{0}$. In fact, according to Theorem 2.3 of the Appendix we know that

$$\mathbf{M}^n\mathbf{u} = \rho^n \begin{Vmatrix} x_1^0 y_1^0 & x_1^0 y_2^0 \\ x_2^0 y_1^0 & x_2^0 y_2^0 \end{Vmatrix} \cdot \mathbf{u} + \mathbf{o}(\rho^n)\mathbf{u},$$

where ρ is the largest eigenvalue of $\mathbf{M}$ and $\mathbf{x}^0 = (x_1^0, x_2^0)$ and $\mathbf{y}^0 = (y_1^0, y_2^0)$ represent the unique (modulo a multiplicative factor) left and right positive eigenvectors normalized so that $x_1^0 y_1^0 + x_2^0 y_2^0 = 1$. The meaning ascribed to the term $\mathbf{o}(\rho^n)$ is an extension of the traditional one. When dividing by ρ^n and letting $n \to \infty$ the quantity $(\mathbf{o}(\rho^n))/\rho^n$ is a matrix each element of which tends to zero. The $\mathbf{o}(\rho^n)$ factor does not depend on $\mathbf{u}$. It represents the error term in the convergence of $\mathbf{M}^n/\rho^n$ to its limit. We rewrite the above expression in the form

$$\mathbf{M}^n\mathbf{u} = \rho^n(y_1^0 s + y_2^0 t)\mathbf{x}^0 + o(\rho^n)\mathbf{u}, \qquad \mathbf{u} = (s,t),$$

and if $\mathbf{u} \geq \mathbf{0}$ we obtain the obvious estimate

$$|\mathbf{M}^n\mathbf{u}| \geq \rho^n[x_1^0 + x_2^0] \min(y_1^0, y_2^0) \cdot |\mathbf{u}| + o(\rho^n)|\mathbf{u}|.$$

Since $\rho > 1$, a sufficiently large choice of n implies (5.10). Combining (5.9) and (5.10) we deduce

$$|\mathbf{1} - \boldsymbol{\varphi}_n(\mathbf{1} - \mathbf{u})| > 2|\mathbf{u}|,$$

provided $\mathbf{1} \geq \mathbf{u} \geq \mathbf{0}$, $|\mathbf{u}|$ is sufficiently small, and n is sufficiently large, say $n \geq n_0$. Let $\mathbf{v} = \mathbf{1} - \mathbf{u}$; then

$$|\mathbf{1} - \boldsymbol{\varphi}_n(\mathbf{v})| > 2|\mathbf{1} - \mathbf{v}| \tag{5.11}$$

for all $\mathbf{0} \leq \mathbf{v} \leq \mathbf{1}$ satisfying $|\mathbf{1} - \mathbf{v}| < \varepsilon$ and $n \geq n_0$. We now utilize (5.11) in order to demonstrate that $\boldsymbol{\pi} \ll 1$. Suppose $\boldsymbol{\pi} = \mathbf{1}$, i.e., $\pi^{(i)} = 1$ for $i = 1, 2$. Then $q_n^{(i)} = \varphi_n^{(i)}(0) \geq 0$ approaches 1 $(n \to \infty)$. Now referring to (5.3) we know that

$$\boldsymbol{\phi}_{n+N}(\mathbf{0}) = \boldsymbol{\phi}_n(\boldsymbol{\phi}_N(\mathbf{0})).$$

Using (5.11) with $\mathbf{v} = \boldsymbol{\phi}_N(\mathbf{0})$ we have

$$|\mathbf{1} - \boldsymbol{\phi}_{n+N}(\mathbf{0})| = |\mathbf{1} - \boldsymbol{\phi}_n(\boldsymbol{\phi}_N(\mathbf{0}))| > 2|\mathbf{1} - \boldsymbol{\phi}_N(\mathbf{0})| \tag{5.12}$$

only if $|\mathbf{1} - \boldsymbol{\phi}_N(\mathbf{0})| < \varepsilon$, and this can be achieved by taking N large enough. However, relation (5.12) contradicts the assumption that $\varphi_n^{(i)}(0)$ tends to 1 as $n \to \infty$. Thus $\pi^{(1)} = \pi^{(2)} = 1$ is impossible. Assume now that $\pi^{(1)} < 1$ and $\pi^{(2)} = 1$. Then

$$\pi^{(1)} = \varphi^{(1)}(\pi^{(1)}, 1)$$

and

$$1 = \pi^{(2)} = \varphi^{(2)}(\pi^{(1)}, 1).$$

Thus, we have

$$\varphi^{(2)}(1, 1) = 1 \qquad \text{and} \qquad \varphi^{(2)}(\pi^{(1)}, 1) = 1,$$

where $\pi^{(1)} < 1$. Since $\varphi^{(2)}(s, t)$ is monotone in s, $\varphi^{(2)}(s, 1)$ must be constant on the interval $\pi^{(1)} \leq s \leq 1$;

$$\frac{\partial \varphi^{(2)}(s, 1)}{\partial s} = 0 \qquad \text{in} \qquad \pi^{(1)} \leq s \leq 1$$

and also

$$m_{21}^{(2)} = \left.\frac{\partial \varphi^{(2)}(s, t)}{\partial s}\right|_{s=t=1} = 0,$$

which clearly contradicts our assumption that $\mathbf{M} \gg \mathbf{0}$. In a similar manner, we deduce that $\pi^{(1)} = 1$, $\pi^{(2)} < 1$ is impossible. Thus $\boldsymbol{\pi} \ll \mathbf{1}$ is established. The verification that $\boldsymbol{\pi}$ is smaller than any other positive fixed point proceeds as follows. Let $\boldsymbol{\pi}^* > \mathbf{0}$ satisfy $\boldsymbol{\phi}(\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$. By monotonicity, we have $\boldsymbol{\pi}^* = \boldsymbol{\phi}(\boldsymbol{\pi}^*) \geq \boldsymbol{\phi}_1(0, 0)$. Iteration produces $\boldsymbol{\pi}^* \geq \boldsymbol{\phi}_n(0, 0)$ and passing to the limit leads to the desired result: $\boldsymbol{\pi}^* \geq \boldsymbol{\pi}$. ∎

We can strengthen the result of Theorem 5.1 and prove

Theorem 5.2. *Under the assumptions of Theorem* 5.1, *if* $\mathbf{q}$ *is any vector in the unit square other than* $\mathbf{1}$ *then* $\lim_{n\to\infty} \boldsymbol{\phi}_n(\mathbf{q}) = \boldsymbol{\pi}$.

Proof. Suppose first that $0 \leq q^i < 1$ $(i = 1, 2)$. If N is a positive integer then the Taylor expansion of $\varphi_n^{(1)}(\mathbf{q})$ has the form

$$\begin{aligned}\varphi_n^{(1)}(\mathbf{q}) &= \Pr\{|\mathbf{X}_n| = 0 | U_0 = 1,\ V_0 = 0\} \\ &\quad + \sum_{0<|\mathbf{x}|\leq N} \Pr\{\mathbf{X}_n = \mathbf{x} | U_0 = 1,\ V_0 = 0\}(q^1)^{x_1}(q^2)^{x_2} \\ &\quad + \sum_{|\mathbf{x}|>N} \Pr\{\mathbf{X}_n = \mathbf{x} | U_0 = 1,\ V_0 = 0\}(q^1)^{x_1}(q^2)^{x_2}.\end{aligned}$$

The last sum is bounded by $(\max(q^1, q^2)^N \Pr\{|\mathbf{X}_n| > N\} \leq (\max(q^1, q^2))^N$ and as $N \to \infty$ this quantity goes to zero since $\max(q^1, q^2) < 1$.

Each coefficient of the first sum goes to zero when $n \to \infty$ since $|\mathbf{X}_n|$ approaches either 0 or ∞. This fact rests on the property that all finite nonzero states are transient. It follows that as $n \to \infty$ with N fixed the first sum tends to zero. This argument proves that

$$\begin{aligned}\lim_{n\to\infty} \varphi_n^{(1)}(\mathbf{q}) &= \lim_{n\to\infty} \Pr\{|\mathbf{X}_n| = 0 | U_0 = 1,\ V_0 = 0\} \\ &= \lim_{n\to\infty} \varphi_n^{(1)}(\mathbf{0}) = \pi^{(1)}\end{aligned}$$

as asserted in the theorem. Similarly

$$\lim_{n\to\infty} \varphi_n^{(2)}(\mathbf{q}) = \pi^{(2)}.$$

If one of the $q^{(i)}$ $(i = 1, 2)$ is equal to 1 but not both, then $\boldsymbol{\phi}_1(\mathbf{q}) = (\varphi^1(\mathbf{q}), \varphi^2(\mathbf{q}))$ determines a nonnegative vector with each component strictly smaller than 1. We apply the preceding analysis to $\boldsymbol{\phi}_1(\mathbf{q})$ and deduce that

$$\lim_{n\to\infty} \boldsymbol{\phi}_n(\boldsymbol{\phi}_1(\mathbf{q})) = \boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\phi}_{n+1}(\mathbf{q}). \quad \blacksquare$$

Corollary 5.1. *The only nonnegative solutions of* (5.5) *are* 1 *and* π.

6: Multi-Type Branching Processes

The generalization of the theory of the preceding section to the case of more than two types proceeds *mutatis mutandis* as in the case of two types. The proofs involve no new ideas or techniques. We merely list the results. The industrious student should supply the detailed proofs.

We will consider a branching growth process consisting of p types. The different types may correspond to actual different mutant forms of an organism or may refer to a single organism where the type distinguishes age or some other like property. The restriction to a finite number of types has the interpretation, for example, that we have specified a finite set of age classifications.

In the case of the production of photons arising in cosmic ray cascades of electrons the type may represent the energy level associated with a photon.

Associated with type i is the probability generating function

$$f^{(i)}(s_1, \ldots, s_p) = \sum_{r_1,\ldots,r_p=0}^{\infty} p^{(i)}(r_1, \ldots, r_p) s_1^{r_1} \cdots s_p^{r_p}, \qquad |s_1| \leq 1, \ldots, |s_p| \leq 1, \quad i = 1, \ldots, p,$$

where $p^{(i)}(r_1, \ldots, r_p)$ is the probability that a single object of type i has r_1 children of type 1, r_2 children of type 2, ..., r_p of type p.

We introduce the vector notation $\mathbf{s} = (s_1, \ldots, s_p)$.

Let $f_n^{(i)}(\mathbf{s})$ denote the nth-generation probability generating function arising from one individual of type i. Analogous to (5.3) we have

$$f_{n+1}^{(i)}(\mathbf{s}) = f^{(i)}(f_n^{(1)}(\mathbf{s}), f_n^{(2)}(\mathbf{s}), \ldots, f_n^{(p)}(\mathbf{s})), \qquad f_0^{(i)}(\mathbf{s}) = s_i, \quad n = 0, 1, \ldots, \quad i = 1, \ldots, p.$$

Let $\mathbf{Z}_n = (Z_n^{(1)}, \ldots, Z_n^{(p)})$ denote the vector representing the population size of p types in the nth generation. The analog of (5.4) is

$$E(\mathbf{Z}_{n+m}|\mathbf{Z}_n) = \mathbf{Z}_n \mathbf{M}^m,$$

where $\mathbf{M} = \|m_{ij}\|_{i,j=1}^{p}$ is the matrix of first moments:

$$m_{ij} = E(Z_1^{(j)}|\mathbf{Z}_0 = \mathbf{e}_i) = \frac{\partial f^{(i)}}{\partial s_j}(1, 1, \ldots, 1), \qquad i, j = 1, \ldots, p,$$

and $\mathbf{e}_i$ denotes the vector with 1 for the ith component and zero otherwise.

We now state the analog of Theorem 5.1 for p types. We will assume $m_{ij} > 0$ for all i, j. (It suffices to have $m_{ij}^{(n)} > 0$ for some n and all i, j.) Let $\pi^{(i)}$ be the extinction probability if initially there is one object of type i ($i = 1, \ldots, p$); that is,

$$\pi^{(i)} = \Pr\{\mathbf{Z}_n = \mathbf{0} \quad \text{for some} \quad n|\mathbf{Z}_0 = \mathbf{e}_i\}.$$

The vector $(\pi^1, \ldots, \pi^p)$ is denoted by $\boldsymbol{\pi}$. Let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1)$.

Theorem 6.1 *Let $m_{ij} > 0$ for all $i, j = 1, \ldots, p$ and let ρ denote the eigenvalue of largest absolute value of the matrix $\mathbf{M}$. If $\rho \leq 1$ then $\boldsymbol{\pi} = \mathbf{1}$. If $\rho > 1$ then $\mathbf{0} \leq \boldsymbol{\pi} \ll \mathbf{1}$ and $\boldsymbol{\pi}$ satisfies the equation*

$$\pi^{(i)} = f^{(i)}(\boldsymbol{\pi}), \qquad i = 1, \ldots, p.$$

7: *Continuous Time Branching Processes*

The branching processes dealt with in Sections 1–6 are limited in that generation times are fixed. Although some phenomena, particularly experimental trials, fit this situation, most natural reproductive processes occur

continuously in time. It is therefore of interest to formulate a continuous time version of branching processes.

In the present section we explore the structure of *time-homogeneous Markov branching processes*; in Section 11 the Markov restriction will be dropped. We determine a continuous time Markov branching process with state variable $X(t) = \{\text{number of particles at time } t, \text{ given } X(0) = 1\}$ by specifying the infinitesimal probabilities of the process. Let

$$\delta_{1k} + a_k h + o(h), \qquad k = 0, 1, 2, \ldots, \tag{7.1}$$

(see Section 4 of Chapter 4 and Chapter 14 of Volume II) represent the probability that a single particle will split producing k particles (or objects) during a small time interval $(t, t+h)$ of length h. In (7.1) δ_{1k} denotes, as customary, the Kronecker delta symbol, and we assume that $a_1 \leq 0$, $a_k \geq 0$ for $k = 0, 2, 3, \ldots$, and

$$\sum_{k=0}^{\infty} a_k = 0. \tag{7.2}$$

We further postulate that individual particles act independently of each other, always governed by the infinitesimal probabilities (7.1). Note that we are effectively assuming time homogeneity for the transition probabilities since a_k is not a function of the time at which the conversion or splitting occurs.

Another way to express the infinitesimal transitions is to differentiate between the time until a split occurs and the nature of the split. Thus each object lives a random length of time following an exponential distribution with mean $\lambda^{-1} = a_0 + a_2 + a_3 + \cdots$. On completion of its lifetime, it produces a random number D of descendants of like objects, where the probability distribution of D is

$$\Pr\{D = k\} = \frac{a_k}{a_0 + a_2 + a_3 + \cdots}, \qquad k = 0, 2, 3, \ldots.$$

The lifetime and progeny distribution of separate individuals are independent and identically distributed. Taking account of the independence assumptions, particularly the property that individuals act independently, we can write (7.1) equivalently in terms of the infinitesimal transition probability matrix as

$$\Pr\{X(t+h) = n + k - 1 | X(t) = n\} = n a_k h + o(h) \tag{7.3}$$

(since in a small time interval one particle on the average will split) and

$$\Pr\{X(t+h) = n | X(t) = n\} = 1 + n a_1 h + o(h), \tag{7.4}$$

where $o(h)/h$ tends to zero as $h \to 0+$.

We have already encountered an example of a continuous time branching process in the guise of a birth and death process. In fact, if we put $a_2 = \lambda$, $a_0 = \mu$, $a_1 = -(\lambda + \mu)$, and $a_k = 0$ otherwise, then $(\lambda + \mu)^{-1}$ can be interpreted as the probability of a birth or death event; $\lambda/(\lambda + \mu)$ $((\mu/(\lambda + \mu)))$ is the probability of a birth (death) under the condition that an event has happened. The stochastic process so obtained whose state variable is population size can now be recognized as a linear growth birth and death process (see Chapter 4, Section 6).

As explained in Chapter 14, it is not a simple matter to construct a Markov process corresponding to a prescribed matrix of infinitesimal probabilities. It is an even more recondite task to assure that the constructed process possesses realizations conforming to the laws of a branching process, i.e., individual particles generate independent families and the descendents act independently, etc. We do not enter the analysis of this construction as it is beyond the scope of this text. The more advanced reader can consult Harris on this point (see the references at the close of this chapter). We further direct attention to Chapter 14 of Volume II for additional discussion on the relations between Markov processes and matrices of infinitesimal probabilities.

Let $P_{ij}(t)$, assumed henceforth well defined, denote the probability that the population of size i at time zero will be of size j at time t, or in symbols $P_{ij}(t) = \Pr\{X(t + s) = j | X(s) = i\}$. As the notation indicates, this probability depends only on the elapsed time, i.e., the process has stationary transition probabilities. We introduce the generating function

$$\phi(t; s) = \sum_{j=0}^{\infty} P_{1j}(t)s^j. \tag{7.5}$$

Since individuals act independently, we have the fundamental relation (cf. page 289)

$$\sum_{j=0}^{\infty} P_{ij}(t)s^j = [\phi(t; s)]^i. \tag{7.6}$$

The formula (7.6) characterizes and distinguishes branching processes from other continuous time Markov chains. It expresses the property that different individuals (i.e., particles) give rise to independent realizations of the process uninfluenced by the pedigrees evolving owing to the other individuals present. In other words the population $X(t; i)$ evolving in time t from i initial parents is the same, probabilistically, as the combined sum of i populations each with one initial parent.

In view of the time homogeneity, the Chapman–Kolmogorov equations take the form

$$P_{ij}(t + \tau) = \sum_{k=0}^{\infty} P_{ik}(t)P_{kj}(\tau). \tag{7.7}$$

With the aid of (7.5), (7.6), and (7.7), we obtain

$$[\phi(t+\tau; s)]^i = \sum_{j=0}^{\infty} P_{ij}(t+\tau)s^j = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} P_{ik}(t)P_{kj}(\tau)s^j$$
$$= \sum_{k=0}^{\infty} P_{ik}(t) \sum_{j=0}^{\infty} P_{kj}(\tau)s^j = \sum_{k=0}^{\infty} P_{ik}(t)[\phi(\tau; s)]^k$$
$$= [\phi(t; \phi(\tau; s))]^i,$$

and in particular

$$\phi(t+\tau; s) = \phi(t; \phi(\tau; s)). \tag{7.8}$$

The relation (7.8) is the continuous time analog of the functional iteration formula of Section 2, fundamental in the case of discrete time branching processes. Next we introduce the generating function of the infinitesimal probabilities defined in (7.1). Specifically, let

$$u(s) = \sum_{k=0}^{\infty} a_k s^k.$$

The following analysis is formal. Consider

$$\phi(h; s) = \sum_{j=0}^{\infty} P_{1j}(h)s^j = \sum_{j=0}^{\infty} (\delta_{1j} + a_j h + o(h))s^j \tag{7.9}$$
$$= s + h \sum_{j=0}^{\infty} a_j s^j + o(h) = s + hu(s) + o(h).$$

From (7.8) with $\tau = h$

$$\phi(t+h; s) = \phi(t; \phi(h; s)) = \phi(t; s + hu(s) + o(h))$$

and expanding the right-hand side with respect to the second variable, by Taylor's theorem, yields

$$\phi(t+h; s) = \phi(t; s) + \frac{\partial\phi(t; s)}{\partial s} hu(s) + o(h).$$

Then

$$\frac{\phi(t+h; s) - \phi(t; s)}{h} = \frac{\partial\phi(t; s)}{\partial s} u(s) + \frac{o(h)}{h}.$$

Letting h decrease to 0 leads to

$$\frac{\partial\phi(t; s)}{\partial t} = \frac{\partial\phi(t; s)}{\partial s} u(s). \tag{7.10}$$

This is a partial differential equation for the function of two variables $\phi(t; s)$, subject to the initial condition

$$\phi(0; s) \equiv \sum_{j=0}^{\infty} P_{1j}(0)s^j \equiv s. \tag{7.11}$$

When $u(s)$ is known, the partial differential equation (7.10) in the presence of (7.11) can be solved for $\phi(t; s)$.

The differential equation (7.10) is merely a form of the forward Kolmogorov differential equations which has been converted into an equivalent differential equation satisfied by the generating function of the transition probability function.

We may derive a second differential equation satisfied by ϕ corresponding to the backward Kolmogorov differential equation. To this end, we substitute $t = h$ in (7.8), which becomes

$$\phi(h + \tau; s) = \phi(h; \phi(\tau; s)),$$

and then use (7.9) with Taylor's expansion as before. This gives

$$\phi(h + \tau; s) = \phi(\tau; s) + hu(\phi(\tau; s)) + o(h).$$

This expression can be written more suggestively as

$$\frac{\phi(\tau + h; s) - \phi(\tau; s)}{h} = u(\phi(\tau; s)) + \frac{o(h)}{h}. \tag{7.12}$$

Letting $h \to 0+$ and replacing τ by t, we obtain

$$\frac{\partial \phi(t; s)}{\partial t} = u(\phi(t; s)). \tag{7.13}$$

This is an ordinary differential equation. The initial condition is again (7.11). Later on we will show how to effectively solve (7.13).

8: *Extinction Probabilities for Continuous Time Branching Processes*

We first carry out the easier task of computing the mean of $X(t)$. To this end, differentiate (7.10) with respect to s and interchange the order of differentiation on the left side. The result is

$$\frac{\partial}{\partial t} \frac{\partial \phi(t; s)}{\partial s} = \frac{\partial^2 \phi(t; s)}{\partial s^2} u(s) + \frac{\partial \phi(t; s)}{\partial s} u'(s). \tag{8.1}$$

Set $s = 1$. Then, since $u(1) = 0$ [Condition (7.2)], we have

$$\frac{\partial m(t)}{\partial t} = u'(1)m(t), \tag{8.2}$$

where

$$m(t) = EX(t) = \left.\frac{\partial \phi(t; s)}{\partial s}\right|_{s=1}.$$

The solution of (8.2) is

$$m(t) = \exp[u'(1)t], \tag{8.3}$$

since the initial condition is $m(0) = 1$ if we assume $X(0) \equiv 1$.

Next we deal with the problem of extinction. In this connection, we assume for the remainder of this section that $a_0 > 0$, as otherwise extinction is impossible. It is enough to consider the case where we start with a single individual at time zero. In fact, from (7.6) we know that

$$\sum_{j=0}^{\infty} P_{ij}(t)s^j = \left[\sum_{j=0}^{\infty} P_{1j}(t)s^j\right]^i.$$

Hence,

$$P_{i0}(t) = [P_{10}(t)]^i.$$

But $P_{i0}(t)$ is the probability of a population of size i dying out by time t. By intuitive considerations we can infer that $P_{i0}(t)$ is nondecreasing in t. We prove this formally by using (7.8). Indeed,

$$P_{i0}(t+\tau) = [\phi(t+\tau; 0)]^i = [\phi(t; \phi(\tau, 0))]^i \geq [\phi(t, 0)]^i = P_{i0}(t),$$

where we used the fact that $\phi(t, s)$ is a power series in s with nonnegative coefficients and is, therefore, an increasing function of s.

The extinction probability may be defined as the probability that the "family" originating from a single individual will eventually die out, i.e.,

$$q = \lim_{t\to\infty} P_{10}(t).$$

Utilizing the theory of discrete time branching processes (Section 3) we can easily determine the probability of extinction in the continuous case. Let t_0 be any fixed positive number and consider the discrete time process

$$X(0), \quad X(t_0), \quad X(2t_0), \ldots, \quad X(nt_0), \ldots,$$

where $X(t)$ is the population size at time t corresponding to the original continuous time branching process that starts with a single individual at time $t = 0$. Since $X(t)$ was assumed to be a Markov process, the discrete time process $Y_n = X(nt_0)$ will obviously be a Markov chain. Moreover, it describes a discrete time branching process. Indeed, by the hypothesis of homogeneity of the probability function of $X(t)$ and by virtue of (7.6), we obtain

$$\begin{aligned}\sum_{k=0}^{\infty} \Pr\{Y_{n+1} = k | Y_n = i\}s^k &= E[s^{Y_{n+1}} | Y_n = i] \\ &= E[s^{X((n+1)t_0)} | X(nt_0) = i] = E[s^{X(t_0)} | X(0) = i] \\ &= [\phi(t_0; s)]^i = \{E[s^{X(t_0)} | X(0) = 1]\}^i \\ &= \{E[s^{Y_1} | Y_0 = 1]\}^i.\end{aligned}$$

This shows that Y_n constitutes a branching process. The generating function of the number of offspring of a single individual in this process is $\phi(t_0; s)$. Hence, we know that the probability of extinction for the Y_n process is the smallest nonnegative root of the equation

$$\phi(t_0; s) = s, \tag{8.4}$$

as was proved in Section 3. But

$$\begin{aligned}\Pr\{Y_n = 0 \quad \text{for some} \quad n\} &= \lim_{n\to\infty} \Pr\{Y_n = 0\} \\ &= \lim_{n\to\infty} \Pr\{X(nt_0) = 0\} \\ &= \lim_{t\to\infty} \Pr\{X(t) = 0\} = q.\end{aligned}$$

Hence, *the extinction probability q of the continuous time branching process $X(t)$ is the smallest nonnegative root of Eq.* (8.4), *where t_0 is any positive number.*

Since q is a root of Eq. (8.4) for any t_0, we expect that we should also be able to calculate q from an equation that does not depend on time. This is indeed the case and we assert the following theorem.

Theorem 8.1. *The probability of extinction q is the smallest nonnegative root of the equation*

$$u(s) = 0. \tag{8.5}$$

Hence, $q = 1$ if and only if $u'(1) \leq 0$. (Recall that $u(s) = \sum_{k=0}^{\infty} a_k s^k = a_1 s + [a_0 + a_2 s^2 + \cdots] = a_1 s + g(s)$.)

Proof. Since q satisfies (8.4) for any t_0, we see on the basis of Eq. (7.12) that

$$0 = u(q) + \frac{o(h)}{h} \qquad \text{for any} \quad h > 0.$$

Letting $h \to 0+$, we obtain $u(q) = 0$.

Since $u''(s) = \sum_{k=2}^{\infty} a_k k(k-1)s^{k-2} \geq 0$, $u(s)$ is convex in the interval $[0, 1]$. As $u(1) = 0$ and $u(0) = a_0 > 0$, $u(s)$ may have at most one zero in $(0, 1)$. According to whether $u'(1) \leq 0$ or $u'(1) > 0$ holds, we have the case represented by Fig. 3 or 4. Notice that $E(X(t_0)) = E(Y) > 1$ if and only if $u'(1) > 0$. This means that for the discrete time branching process $X(nt_0)$, $n = 0, 1, 2, \ldots$ ($t_0 > 0$ fixed), extinction occurs with probability < 1 and therefore the same is true for the process $X(t)$. The probability of extinction q is in this case necessarily the smaller zero of $u(s)$ in $[0, 1]$. In a similar manner we conclude that if $u'(1) \leq 0$, q must equal one. In either case q is the smallest nonnegative root of (8.5).

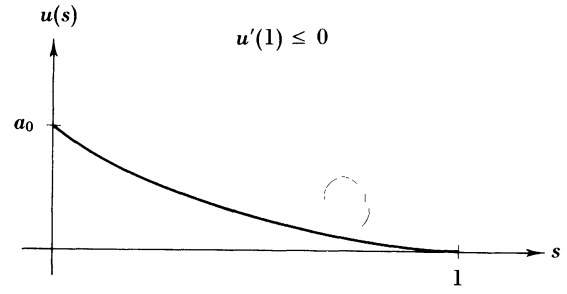


FIG. 3

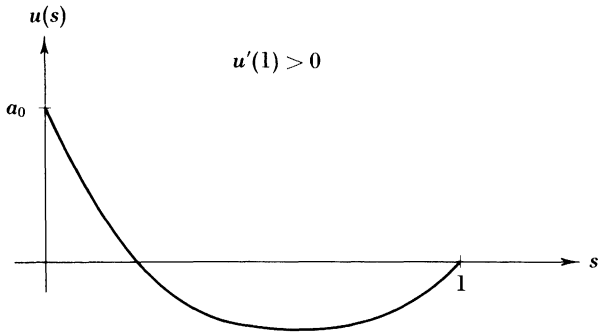


FIG. 4

9: *Limit Theorems for Continuous Time Branching Processes*

We turn to the task of solving the ordinary differential equation (7.13) and analyzing and interpreting its growth properties as $t \to \infty$. Since $\exp[u'(1)t]$ is the expected number of particles at time t we anticipate different behavior according to whether $u'(1)$ is negative, zero, or positive. We will only discuss the case when $u'(1) < 0$ under the additional assumption that $u''(1) < \infty$. First we prove that the function

$$B(s) = \frac{1}{u(s)} - \frac{1}{u'(1)(s-1)}$$

is bounded and hence integrable in $0 \leq s < 1$. Indeed, expanding $u(s)$ in the neighborhood of $s = 1$ leads to the formula

$$u(s) = u(1) + u'(1)(s-1) + R(s)(s-1)^2, \qquad s \leq 1,$$

where

$$\lim_{s \to 1-} R(s) = \frac{u''(1)}{2!} < \infty. \tag{9.1}$$

Then, recalling that $u(q) = u(1) = 0$, we have

$$\frac{1}{u(s)} = \frac{1}{u'(1)(s-1) + R(s)(s-1)^2} = \frac{1}{u'(1)(s-1)} \cdot \frac{1}{1 + [R(s)(s-1)/u'(1)]}$$

$$= \frac{1}{u'(1)(s-1)} \left\{ 1 - \frac{R(s)(s-1)/u'(1)}{1 + [R(s)(s-1)/u'(1)]} \right\}.$$

Hence,

$$B(s) = -\frac{R(s)/[u'(1)]^2}{1 + [R(s)(s-1)/u'(1)]} \tag{9.2}$$

and we can now infer as a direct consequence of (9.1) that $B(s)$ is bounded in the neighborhood of $s = 1-$. Certainly $B(s)$ is bounded for s away from $s = 1$, i.e., for $0 \leq s \leq 1 - \delta$, as is evident from its definition, since $u(s)$ vanishes only at $s = 1$ in the case under consideration $[u'(1) < 0]$. Thus, $B(s)$ is bounded in $0 \leq s < 1$ under the conditions $u''(1) < \infty$, and $u'(1) < 0$. Now we may define, for $0 \leq s < 1$,

$$K(s) = \int_1^s \left[\frac{1}{u(x)} - \frac{1}{u'(1)(x-1)} \right] dx + \frac{\log(1-s)}{u'(1)}, \tag{9.3}$$

as the integral exists and is finite.
Notice, further, that

$$K'(s) = \frac{1}{u(s)} > 0 \qquad \text{for} \quad 0 \leq s < 1,$$

again owing to the assumption $u'(1) < 0$. This means that $K(s)$ is strictly increasing and continuous; hence, the mapping

$$w = K(s) \tag{9.4}$$

possesses a continuous strictly increasing inverse function

$$s = K^{-1}(w) = L(w), \qquad L(K(s)) = s, \tag{9.5}$$

with the property that as s traverses $[0, 1)$, w traverses $[K(0), \infty)$, and observe that $K(0) < 0$. We are now in possession of the ingredients needed to exhibit the desired solution of (7.13) under the initial condition (7.11). Separation of variables in (7.13) and integration lead to an implicit formula for $\phi(t, s)$:

$$\int_s^{\phi(t;s)} \frac{dx}{u(x)} = t.$$

Performing obvious rearrangements and using the definition of $K(\cdot)$, we obtain

$$t = \int_s^{\phi(t;s)} \frac{dx}{u(x)} = \int_s^{\phi(t;s)} \left[\frac{1}{u(x)} - \frac{1}{u'(1)(x-1)}\right] dx + \frac{1}{u'(1)} \log(1-x)\Big|_s^{\phi(t;s)}$$

$$= \int_1^{\phi(t;s)} \left[\frac{1}{u(x)} - \frac{1}{u'(1)(x-1)}\right] dx + \frac{\log(1-\phi(t;s))}{u'(1)}$$

$$- \int_1^{s} \left[\frac{1}{u(x)} - \frac{1}{u'(1)(x-1)}\right] dx - \frac{\log(1-s)}{u'(1)}$$

$$= K(\phi(t;s)) - K(s).$$

Equivalently, we have the relation

$$K(\phi(t;s)) = t + K(s).$$

Since the inverse function exists, this becomes

$$\phi(t;s) = K^{-1}(t + K(s)) \qquad \text{for} \quad 0 \le s < 1 \quad \text{and} \quad t \ge 0. \tag{9.6}$$

Under the assumptions $u'(1) < 0$, $u''(1) < \infty$ we may also derive some asymptotic results for the probability of extinction in time t $(t \to \infty)$. Because $B(s)$ is bounded in $0 \le s < 1$ and $\lim_{s \to 1-} B(s)$ exists we may write [see (9.3)]

$$K(s) = \frac{\log(1-s)}{u'(1)} - C \cdot (1-s) + o(1-s) \tag{9.7}$$

in the neighborhood of $s = 1$. Here C is a negative constant; in fact, $\lim_{x \to 1-} [1/u(x) - (1/(x-1)u'(1))] = C = (-u''(1)/2[u'(1)]^2)$. The $o(1-s)$ term has the usual interpretation, i.e., $[o(1-s)]/(1-s) \to 0$ as $s \to 1-$. Rearranging the last expression in the form

$$\log(1-s) = u'(1)K(s) - Cu'(1)(1-s) + o(1-s)$$

and taking exponentials yields

$$1 - s = \exp[u'(1)K(s)] \exp[-Cu'(1)(1-s)] \exp[o(1-s)].$$

But

$$\exp[o(1-s)] = 1 + o(1-s)$$

and

$$\exp[-Cu'(1)(1-s)] = 1 - Cu'(1)(1-s) + o(1-s).$$

It follows that

$$1 - s = \exp[u'(1)K(s)][1 - Cu'(1)(1-s) + o(1-s)]. \tag{9.8}$$

Consequently

$$\lim_{s\to 1-} \frac{1-s}{\exp[u'(1)K(s)]} = 1.$$

By virtue of this limit relation we can write (9.8) in the form

$$1-s = \{\exp[u'(1)K(s)]\}(1 - Cu'(1)\exp[u'(1)K(s)] + o(\exp[u'(1)K(s)]).$$

Replacing s throughout by $K^{-1}(w)$ [see (9.5)] we get

$$1 - K^{-1}(w) = \exp[u'(1)w][1 - Cu'(1)\exp[u'(1)w] + o(\exp[u'(1)w])\}, \tag{9.9}$$

where $s\to 1-$ is equivalent to $w\to\infty$. Now with the aid of (9.6) and (9.9), we can calculate the probability of no extinction in time t. Thus

$$\begin{aligned} 1 - P_{10}(t) &= 1 - \phi(t;0) = 1 - K^{-1}(t + K(0)) \\ &= [\exp\{u'(1)(t + K(0))\}] \\ &\quad \times \{1 - Cu'(1)(\exp[u'(1)(t + K(0))] + o(\exp[u'(1)(t + K(0))])\}] \\ &= \exp\{u'(1)K(0)\}\exp\{u'(1)t\} + O(\exp\{2u'(1)t\}) + o(\exp[u'(1)t]) \end{aligned}$$

or, equivalently,

$$1 - P_{10}(t) = \exp[u'(1)K(0)]m(t) + o(\exp[u'(1)t]). \tag{9.10}$$

Another asymptotic result, as $t\to\infty$, may be obtained as follows. The conditional probability generating function of $X(t)$, given that $X(t)\neq 0$, is defined as

$$g(z,t) = \sum_{k=0}^{\infty} \Pr\{X(t) = k | X(t)\neq 0\}z^k, \qquad 0\le z < 1.$$

But

$$\begin{aligned} \Pr\{X(t) = k | X(t)\neq 0\} &= \frac{\Pr\{X(t) = k,\ X(t)\neq 0\}}{\Pr\{X(t)\neq 0\}} \\ &= \begin{cases} 0 & \text{if } k = 0, \\ \dfrac{\Pr\{X(t) = k\}}{1 - \Pr\{X(t) = 0\}} & \text{if } k\neq 0. \end{cases} \end{aligned}$$

Thus

$$\begin{aligned} g(z;t) &= \sum_{k=1}^{\infty} \frac{\Pr\{X(t) = k\}}{1 - \Pr\{X(t) = 0\}} z^k = \frac{\phi(t;z) - \phi(t;0)}{1 - \phi(t;0)} \\ &= \frac{K^{-1}(t + K(z)) - K^{-1}(t + K(0))}{1 - K^{-1}(t + K(0))} \\ &= \frac{[1 - K^{-1}(t + K(0))] - [1 - K^{-1}(t + K(z))]}{1 - K^{-1}(t + K(0))} \end{aligned}$$

where formula (9.6) is used for $\phi(t;s)$.

Substituting the expression for $K^{-1}(w)$ given by (9.9) yields

$$g(z;t) = \frac{e^{u'(1)(t+K(0))}[1+O(e^{u'(1)(t+K(0))})] - e^{u'(1)(t+K(z))}[1+O(e^{u'(1)(t+K(z))})]}{e^{u'(1)(t+K(0))}[1+O(e^{u'(1)(t+K(0))})]}$$

$$= 1 - e^{u'(1)[K(z)-K(0)]}\frac{1+O(e^{u'(1)(t+K(z))})}{1+O(e^{u'(1)(t+K(0))})}.$$

Now let $t \to \infty$. Then the fraction on the right-hand side reduces to 1 by the assumption $u'(1) < 0$. Hence

$$\lim_{t\to\infty} g(z;t) = g(z) = 1 - \exp\{u'(1)[K(z) - K(0)]\}.$$

By (9.3), however,

$$K(z) - K(0) = \int_0^z \left[\frac{1}{u(x)} - \frac{1}{u'(1)(x-1)}\right] dx + \frac{\log(1-z)}{u'(1)}$$

$$= \int_0^z \frac{dx}{u(x)} - \frac{\log(1-x)}{u'(1)}\Big|_0^z + \frac{\log(1-z)}{u'(1)} = \int_0^z \frac{dx}{u(x)}.$$

Hence, as $t \to \infty$ we have the limit probability generating function

$$g(z) = 1 - \exp\left[u'(1)\int_0^z \frac{dx}{u(x)}\right] = \sum_{k=1}^{\infty} \lim_{t\to\infty} \Pr\{X(t) = k | X(t) \neq 0\} z^k.$$

We summarize the preceding discussion in the statement of the following theorem:

Theorem 9.1. *Consider a continuous time branching process $X(t)$ determined by the infinitesimal generating function*

$$u(s) = \sum_{k=0}^{\infty} a_k s^k, \tag{9.11}$$

where the $\{a_k\}$ possesses the interpretation given in (7.1) *and is subject to the conditions* (7.2). *Suppose that $u''(1) < \infty$. Suppose further that $u'(1) < 0$ so that the extinction probability $q = 1$* (see Theorem 8.1). *Then*

$$\phi(t,s) = \sum_{k=0}^{\infty} \Pr\{X(t) = k | X(0) = 1\} s^k \tag{9.12}$$

$$= K^{-1}(t + K(s)), \qquad t \geq 0 \qquad |s| < 1,$$

where $K(s)$ is defined in (9.3). *The probability of no extinction by time t tends to zero at an exponential rate according to*

$$\lim_{t\to\infty} \frac{1 - P_{10}(t)}{\exp[u'(1)K(0)]\exp[u'(1)t]} = 1.$$

Moreover, the random variable $X(t)$ conditioned by $X(t) > 0$ has a limit distribution whose probability generating function is given by

$$g(z, t) = \frac{\sum_{k=1}^{\infty} \Pr\{X(t) = k | X(0) = 1\} z^k}{1 - \Pr\{X(t) = 0\}}$$

$$\to 1 - \exp\left[u'(1) \int_0^z \frac{dx}{u(x)}\right] \quad \text{as} \quad t \to \infty. \tag{9.13}$$

We state without proof the following limit theorem which corresponds to the case $u'(1) = 0$ and $u'(1) > 0$. Their proofs are more complicated although similar in substance.

Theorem 9.2. (i) *Suppose $u'(1) = 0$, $u''(1) < \infty$. Then*

$$\Pr\{X(t) > 0 | X(0) = 1\} \sim \frac{2}{u''(t)} \frac{1}{t}, \qquad t \to \infty,$$

and

$$\lim_{t \to \infty} \Pr\left\{\frac{2X(t)}{u''(1)t} > \lambda | X(t) > 0\right\} = e^{-\lambda}, \qquad \lambda > 0.$$

(ii) *If $u'(1) > 0$ and $u''(1) < \infty$, then*

$$Z(t) = \frac{X(t)}{\exp[u'(1)t]}$$

has a limit distribution as $t \to \infty$.

10: Two-Type Continuous Time Branching Process

Consider two different types of particles which we will call type 1 and type 2 particles, respectively. A continuous time branching Markov process for two types of particles will be determined by appropriately specifying the infinitesimal probabilities [(7.3) and (7.4)]. Explicitly, we postulate that each particle of type $i (i = 1, 2)$ may at any time, independent of its past and independent of the history or present state of any of the other particles of either type, convert during a small time interval $(t, t + h)$ into k_1 and k_2 particles of types 1 and 2, respectively, with probabilities

$$\delta_{1k_1}\delta_{0k_2} + a^{(1)}_{k_1,k_2} h + o(h)$$

(δ_{ij} denotes the familiar Kronecker delta symbol) for a single parent of type 1 and

$$\delta_{0k_1}\delta_{1k_2} + a^{(2)}_{k_1,k_2} h + o(h)$$

for a single parent of type 2 ($k_1, k_2 = 0, 1, 2, \ldots$). Note that we are again postulating time homogeneity for the transition probabilities in that the constants $a^{(i)}_{k_1,k_2}$ are time independent. The parameters obey the restrictions

$$\begin{aligned} &a^{(1)}_{1,0} \leq 0, && a^{(2)}_{0,1} \leq 0, \\ &a^{(1)}_{k_1,k_2} \geq 0 && \text{for all } k_1, k_2 = 0, 1, 2, \ldots \\ & && \text{except } k_1 = 1, k_2 = 0, \\ &a^{(2)}_{k_1,k_2} \geq 0 && \text{for all } k_1, k_2 = 0, 1, 2, \ldots \\ & && \text{except } k_1 = 0, k_2 = 1, \end{aligned}$$

and

$$\sum_{k_1,k_2=0}^{\infty} a^{(i)}_{k_1,k_2} = 0, \qquad i = 1, 2.$$

We introduce the pair of infinitesimal generating functions

$$u^{(i)}(s_1, s_2) = \sum_{k_1,k_2=0}^{\infty} a^{(i)}_{k_1,k_2} s_1^{k_1} s_2^{k_2}, \qquad i = 1, 2 \qquad (|s_1| \leq 1, |s_2| \leq 1).$$

Let $P_{k_1,k_2;j_1,j_2}(t)$ be the probability that a population of k_1 objects of type 1 and k_2 objects of type 2 present at time 0 will be transformed into a population consisting of j_1 objects of type 1 and j_2 objects of type 2 over a time period of length t. Since the infinitesimal probabilities $a^{(i)}_{k_1,k_2}$ were defined to be independent of time, the transition probabilities are necessarily time homogeneous. We define the probability generating functions

$$\phi^{(1)}(t; s_1, s_2) = \sum_{j_1,j_2=0}^{\infty} P_{1,0;j_1j_2}(t) s_1^{j_1} s_2^{j_2},$$

$$\phi^{(2)}(t; s_1, s_2) = \sum_{j_1,j_2=0}^{\infty} P_{0,1;j_1j_2}(t) s_1^{j_1} s_2^{j_2}.$$

Then it follows analogously as in the model of the one-type branching process that

$$\sum_{j_1,j_2=0}^{\infty} P_{k_1,k_2;j_1j_2}(t) s_1^{j_1} s_2^{j_2} = [\phi^{(1)}(t; s_1, s_2)]^{k_1} [\phi^{(2)}(t; s_1, s_2)]^{k_2} \qquad (k_1, k_2 = 0, 1, 2, \ldots). \tag{10.1}$$

In fact, (10.1) can be regarded as the defining relation of a continuous time two-type branching process. In other words any transition probability matrix function satisfying (10.1) is said to generate a two-type continuous time Markov branching process. The Markov character of the process is summarized in the Chapman–Kolmogorov equations

$$P_{k_1,k_2,j_1,j_2}(t + \tau) = \sum_{l_1,l_2=0}^{\infty} P_{k_1,k_2,;l_1,l_2}(t)\, P_{l_1,l_2;j_1,j_2}(\tau). \tag{10.2}$$

Then from (10.1) and (10.2)

$$\begin{aligned}\phi^{(1)}(t+\tau; s_1, s_2) &= \sum_{j_1,j_2=0}^{\infty} P_{1,0;j_1,j_2}(t+\tau) s_1^{j_1} s_2^{j_2} \\ &= \sum_{j_1,j_2=0}^{\infty} \sum_{l_1,l_2=0}^{\infty} P_{1,0;l_1,l_2}(t) P_{l_1,l_2;j_1,j_2}(\tau)\, s_1^{j_1} s_2^{j_2} \\ &= \sum_{l_1,l_2=0}^{\infty} P_{1,0;l_1,l_2}(t) \sum_{j_1,j_2=0}^{\infty} P_{l_1,l_2;j_1,j_2}(\tau)\, s_1^{j_1} s_2^{j_2} \\ &= \sum_{l_1,l_2=0}^{\infty} P_{1,0;l_1,l_2}(t) [\phi^{(1)}(\tau; s_1, s_2)]^{l_1} [\phi^{(2)}(\tau; s_1, s_2)]^{l_2} \\ &= \phi^{(1)}(t; \phi^{(1)}(\tau; s_1, s_2), \phi^{(2)}(\tau; s_1, s_2)).\end{aligned}$$

The same procedure applied to the generating function $\phi^{(2)}(t; s_1, s_2)$ yields

$$\phi^{(i)}(t+\tau; s_1, s_2) = \phi^{(i)}(t; \phi^{(1)}(\tau; s_1, s_2), \phi^{(2)}(\tau; s_1, s_2)) \quad \text{for } i = 1, 2. \tag{10.3}$$

Moreover,

$$\begin{aligned}\phi^{(1)}(h; s_1, s_2) &= \sum_{j_1,j_2=0}^{\infty} P_{1,0;j_1,j_2}(h)\, s_1^{j_1} s_2^{j_2} \\ &= \sum_{j_1,j_2=0}^{\infty} [\delta_{1j_1}\delta_{0j_2} + a_{j_1j_2}^{(1)} h + o(h)]\, s_1^{j_1} s_2^{j_2} \\ &= s_1 + h u^{(1)}(s_1, s_2) + o(h)\end{aligned}$$

and similarly for $\phi^{(2)}(h; s_1, s_2)$. Thus we have

$$\phi^{(i)}(h; s_1, s_2) = s_i + h u^{(i)}(s_1, s_2) + o(h), \qquad i = 1, 2. \tag{10.4}$$

We now derive a pair of partial differential equations satisfied by $\phi^{(i)}(t; s_1, s_2)$ $(i = 1, 2)$, analogous to (7.10) and (7.13). To this end, we start by setting $\tau = h$ and substituting (10.4) in (10.3). Using Taylor's expansion, we obtain

$$\begin{aligned}&\phi^{(i)}(t+h; s_1, s_2) = \\ &\qquad \phi^{(i)}(t; s_1 + h u^{(1)}(s_1, s_2) + o(h), s_2 + h u^{(2)}(s_1, s_2) + o(h)) \\ &\qquad = \phi^{(i)}(t; s_1, s_2) + \frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial s_1} h u^{(1)}(s_1, s_2) \\ &\qquad\quad + \frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial s_2} h u^{(2)}(s_1, s_2) + o(h)\end{aligned}$$

Dividing both sides by h and letting $h \to 0$ we formally obtain the differential equations

$$\frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial t} = \frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial s_1} u^{(1)}(s_1, s_2) + \frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial s_2} u^{(2)}(s_1, s_2), \qquad i = 1, 2. \tag{10.5}$$

We start again with (10.3), this time setting $t = h$ and using (10.4); this leads to the formula

$$\begin{aligned}\phi^{(i)}(h + \tau; s_1, s_2) &= \phi^{(i)}\,(h; \phi^{(1)}(\tau; s_1, s_2), \phi^{(2)}(\tau; s_1, s_2)) \\ &= \phi^{(i)}(\tau; s_1, s_2) + h u^{(i)}(\phi^{(1)}(\tau; s_1, s_2), \phi^{(2)}(\tau; s_1, s_2)) + o(h).\end{aligned}$$

Then dividing by h, letting $h \to 0$, and finally writing t in place of τ, we obtain a second system of differential equations:

$$\frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial t} = u^{(i)}(\phi^{(1)}(t; s_1, s_2), \phi^{(2)}(t; s_1, s_2)), \; i = 1,2. \tag{10.6}$$

The initial conditions for both (10.5) and (10.6) are

$$\phi^{(i)}(0; s_1, s_2) = s_i, \qquad i = 1, 2.$$

With the aid of (10.5) and (10.6) we can derive systems of ordinary differential equations satisfied by the moments of the random variables of the process. We will not enter into details of these calculations here.

We next offer some applications and examples of two-type continuous time branching processes.

Example 1. Our first example involves a branching process with immigration. We consider the one-type continuous time branching process and enlarge its scope by allowing, in addition to branching, some migration of particles into the system. Recall that

$$\delta_{1k} + a_k h + o(h), \qquad k = 0, 1, 2, \ldots,$$

represents the probability that a particle will convert into k particles during a small time interval $(t, t + h)$ independent of its past and of all other particles. Let us superimpose immigration into the population as follows. Specifically, let

$$\delta_{0k} + b_k h + o(h), \qquad k = 0, 1, 2, \ldots,$$

denote the probability, independent of the present or past history of the population, that k particles of the same kind immigrate and merge with

the population during the time interval $(t, t+h)$. Note that the parameters a_k as well as the parameters b_k are assumed to be independent of the precise time t that the conversion or the immigration takes place. In other words the associated infinitesimal transition probabilities per individual are time homogeneous. For the a_k and b_k we impose the conditions

$$\begin{aligned} &a_1 \leq 0, \qquad b_0 \leq 0, \\ &a_k \geq 0 \qquad \text{for} \quad k = 0, 2, 3, \ldots, \\ &b_k \geq 0 \qquad \text{for} \quad k = 1, 2, 3, \ldots, \\ &\sum_{k=0}^{\infty} a_k = \sum_{k=0}^{\infty} b_k = 0. \end{aligned}$$

Let

$$\begin{aligned} P_k(t) &= \Pr\left\{\begin{matrix}\text{population at time } t \text{ is of size } k \text{ if} \\ \text{there were no particles at time } t = 0\end{matrix}\right\} \\ &= \Pr\{X(t) = k | X(0) = 0\} \qquad k = 0, 1, 2, \ldots, \end{aligned} \tag{10.7}$$

and denote its generating function by

$$\phi(t; s) = \sum_{k=0}^{\infty} P_k(t) s^k. \tag{10.8}$$

Our objective is to evaluate $P_k(t)$ or, if this is not feasible, to ascertain some of its properties.

We introduce the infinitesimal generating functions

$$u(s) = \sum_{k=0}^{\infty} a_k s^k \qquad \text{and} \qquad v(s) = \sum_{k=0}^{\infty} b_k s^k.$$

It is possible to cast the one-type continuous time branching process with immigration in the form of a two-type continuous time branching process. This is done as follows. Assume that we have two different types of particles, types 1 and 2, with infinitesimal probabilities of conversion which we will now specialize, as described at the start of this section.

The idea underlying the identification runs as follows. We have available two types of particles: the first is real, while the second is of a fictitious nature. Real particles upon termination of their lifetime (which is of random duration distributed according to an exponential law with parameter $\lambda^{-1} = a_0 + a_2 + a_3 + \cdots$) create k new real particles with probability $\lambda \cdot a_k$ $(k = 0, 2, 3, \ldots)$. A fictitious particle also lives a random length of time (exponentially distributed with parameter $\bar{\lambda}^{-1} = b_1 + b_2 + \cdots$) and at the end of its life produces l real particles and one further fictitious particle with probability $\bar{\lambda} \cdot b_l$ $(l = 1, 2, 3, \ldots)$. Notice that $\sum_{l=1}^{\infty} \bar{\lambda} b_l = 1$.

The progeny of fictitious particles account for the immigration factor. Thus, we set

$$a^{(1)}_{k_1,k_2} = \begin{cases} a_{k_1} & \text{if } k_2 = 0, \\ 0 & \text{if } k_2 \neq 0, \end{cases}$$

$$a^{(2)}_{k_1,k_2} = \begin{cases} b_{k_1} & \text{if } k_2 = 1, \\ 0 & \text{if } k_2 \neq 1. \end{cases}$$

Then in accordance with the notation of the beginning of this section we have

$$u^{(1)}(s_1, s_2) = \sum_{k_1,k_2=0}^{\infty} a^{(1)}_{k_1,k_2} s_1^{k_1} s_2^{k_2} = \sum_{k_1=0}^{\infty} a_{k_1} s_1^{k_1},$$

$$u^{(2)}(s_1, s_2) = \sum_{k_1,k_2=0}^{\infty} a^{(2)}_{k_1,k_2} s_1^{k_1} s_2^{k_2} = s_2 \sum_{k_1=0}^{\infty} b_{k_1} s_1^{k_1}.$$

Thus

$$u^{(1)}(s_1, s_2) = u(s_1), \qquad u^{(2)}(s_1, s_2) = s_2 v(s_1).$$

In the special case under consideration the differential equation (10.5) reduces to

$$\frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial t} = \frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial s_1} u(s_1) + \frac{\partial \phi^{(i)}(t; s_1, s_2)}{\partial s_2} s_2 v(s_1), \qquad i = 1, 2, \tag{10.9}$$

and the differential equation (10.6) becomes

$$\frac{\partial \phi^{(1)}(t; s_1, s_2)}{\partial t} = u(\phi^{(1)}(t; s_1, s_2)) \tag{10.10}$$

and

$$\frac{\partial \phi^{(2)}(t; s_1, s_2)}{\partial t} = [\phi^{(2)}(t; s_1, s_2)] v(\phi^{(1)}(t; s_1, s_2)). \tag{10.11}$$

Now we will relate the probabilities $P_{0,1;j_1,j_2}(t)$ of the two-type process to the probabilities defined in (10.7). In accordance with the meaning ascribed to the two types of particles the initial state (0, 1) signifies that we start at time 0 with no real particles but with the presence of a potential immigrant signified by a fictitious particle. By the very meaning of the symbols we obviously have

$$P_{0,1;j_1,j_2}(t) = \begin{cases} P_{j_1}(t) & \text{if } j_2 = 1, \\ 0 & \text{if } j_2 \neq 1, \end{cases}$$

and hence

$$\phi^{(2)}(t; s_1, s_2) = s_2 \phi(t; s_1). \tag{10.12}$$

Then from (10.9) we obtain

$$\frac{\partial \phi(t; s)}{\partial t} = \frac{\partial \phi(t; s)}{\partial s} u(s) + \phi(t; s)v(s), \tag{10.13}$$

where we have written s in place of s_1. The initial condition here is

$$\phi(0; s) = 1. \tag{10.14}$$

Instead of solving this differential equation, it is easier to solve the system (10.10) and (10.11) with appropriate initial conditions. Equation (10.10) can be dealt with by methods paraphrasing the analysis of (7.13). The solution of (10.10) can be represented as in (9.6). We denote the solution of (10.10) by $f(t; s)$, where s has taken the place of s_1 and s_2 is suppressed. Because of (10.12), (10.11) becomes

$$\frac{\partial \phi(t; s)}{\partial t} = \phi(t; s)v(f(t; s)), \tag{10.15}$$

with initial condition (10.14). The solution of (10.15) is

$$\phi(t; s) = \exp\left[\int_0^t v(f(\tau; s))\, d\tau\right].$$

Example 2. We close this section by describing a simple, binary-fission, *non*-Markov, one-type, continuous time, branching process that can be reduced to a two-type, continuous time, Markov branching process. Assume that a particle has a lifetime distribution with density

$$\frac{\lambda^2}{2} te^{-\lambda t}, \tag{10.16}$$

i.e., a gamma distribution of order 2. When the particle dies it is replaced by two particles of the same kind, each independent of the other and of the original particle, and each following the lifetime distribution (10.16).

Markov processes are generally characterized by the property that the waiting time in any given state is exponentially distributed. In the present context the waiting time in a given state is determined by the lifetime of the particle. If this is exponential then the population process of these particles constitutes a Markov process. In the growth model introduced above, lifetime does not follow an exponential distribution but that of a convolution of two exponentials.

Let $X(t)$ represent the number of particles at time t and assume that $X(0) = 1$. Since (10.16) is the density of the sum of two independent exponentially distributed r.v's each with parameter λ, we may regard

each particle as going through two separate phases of life each with an exponentially distributed lifetime of parameter λ. This may easily be reduced to the two-type, continuous time, Markov branching process. Instead of referring to two phases of life for the same particle, we will talk about two different types of particles. A particle of type 1 has an exponential distribution lifetime with parameter λ and then converts into a particle of type 2. A particle of type 2 has an exponential distribution lifetime with parameter λ and then converts into two particles of type 1. Thus, conforming to the notation of the beginning of this section we have

$$a^{(1)}_{k_1,k_2} = \begin{cases} -1 & \text{if } k_1 = 1 \text{ and } k_2 = 0, \\ +1 & \text{if } k_1 = 0 \text{ and } k_2 = 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$a^{(2)}_{k_1,k_2} = \begin{cases} +1 & \text{if } k_1 = 2 \text{ and } k_2 = 0, \\ -1 & \text{if } k_1 = 0 \text{ and } k_2 = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where, for simplicity, we have assumed $\lambda = 1$. Then,

$$u^{(1)}(s_1, s_2) = -s_1 + s_2$$

and

$$u^{(2)}(s_1, s_2) = s_1^2 - s_2 .$$

The relations (10.5) and (10.6) will take special forms. The generating function of $X(t)$, given $X(0) = 1$, can be obtained from $\phi^{(1)}(t; s_1, s_2)$ by setting $s_1 = s_2 = s$.

11: Branching Processes with General Variable Lifetime

In this section we will discuss a branching process model where each object (or particle or individual) lives a random length of time following a general lifetime distribution and at the culmination of its life produces its progeny. This process should be compared with branching processes of fixed lifetime or of exponentially distributed lifetime. We assume that an individual object has a lifetime of random length T with probability density function $f(t)$; that is, the probability that this organism will die during the time interval $(t, t + dt)$ is $f(t)\,dt$. We further assume that at the time of its death the object splits into two, thus creating two new objects of like kind. These will have independently distributed random lifetimes with the same density function $f(t)$. At the end of its life each object splits again into two new objects of the same kind and this process continues

indefinitely. Let $N(t)$ denote the number of objects existing at time t, and represent its probability distribution by

$$p_k(t) = \Pr\{N(t) = k\} \qquad \text{for} \quad k = 0, 1, 2, \ldots.$$

Clearly $p_0(t) = 0$ for all $t \geq 0$, as we will always have at least one object. In fact, we will have exactly one before the first split occurs and at least two after the first split. Thus

$$p_1(t) = \Pr\{N(t) = 1\} = \Pr\{T > t\} = 1 - F(t),$$

where

$$F(t) = \Pr\{T \leq t\} = \int_0^t f(\tau)\, d\tau$$

is the cumulative distribution function of T.

Generally, let $G(s, t)$ be the probability generating function of $N(t)$, i.e.,

$$G(s, t) = \sum_{k=0}^{\infty} p_k(t) s^k = \sum_{k=1}^{\infty} p_k(t) s^k.$$

We will now obtain an integral equation for $G(s, t)$. The probability of having exactly k ($k = 2, 3, \ldots$) objects at time t, $p_k(t)$, can be evaluated as follows. Assume that the first fission occurs between time τ and $\tau + d\tau$ ($0 \leq \tau \leq t$) with probability $f(\tau)\, d\tau$, and that each of the two new objects independently undergoing the same branching process will produce a total number k of descendants during the remaining time of length $t - \tau$. Naturally the time, τ, of the first split may take any value in the interval $[0, t]$. Thus, from the law of total probabilities we obtain

$$p_k(t) = \int_0^t d\tau\, f(\tau) \sum_{l=1}^{k} p_l(t - \tau) p_{k-l}(t - \tau), \qquad k = 2, 3, \ldots,$$

and

$$p_1(t) = 1 - F(t).$$

Then from the definition of $G(s, t)$, we have

$$\begin{aligned} G(s, t) &= [1 - F(t)]s + \sum_{k=2}^{\infty} s^k \int_0^t d\tau\, f(\tau) \sum_{l=1}^{k} p_l(t - \tau) p_{k-l}(t - \tau) \\ &= [1 - F(t)]s + \sum_{k=0}^{\infty} s^k \int_0^t d\tau\, f(\tau) \sum_{l=0}^{k} p_l(t - \tau) p_{k-l}(t - \tau). \end{aligned}$$

Since all quantities involved are nonnegative, the summations and the integral sign may be interchanged, and

$$G(s, t) = \int_0^t d\tau\, f(\tau) \sum_{k=0}^{\infty} s^k \sum_{l=0}^{k} p_l(t - \tau) p_{k-l}(t - \tau) + [1 - F(t)]s.$$

We recognize the sum as a convolution, each factor of which is $\sum_{k=0}^{\infty} s^k p_k(t-\tau) = G(s, t-\tau)$. Thus,

$$G(s, t) = \int_0^t [G(s, t-\tau)]^2 f(\tau)\, d\tau + [1 - F(t)]s. \tag{11.1}$$

Unfortunately, this integral equation cannot be solved in general. We will solve it, however, in the special case when T has the exponential distribution with density

$$f(t) = \lambda e^{-\lambda t} \qquad \text{for} \quad t \geq 0. \tag{11.2}$$

The process corresponding to this special case is equivalent to the Yule pure birth process. In fact, if there are n initial objects then the time interval until the first split is the random variable $Z = \min(X_1, X_2, \ldots, X_n)$, where the X_i are independent and possess the distribution law (11.2). The distribution of Z is exponential with parameter $n\lambda$. Therefore, the chance of a split during the next h units of time is $n\lambda h + o(h)$. When this occurs the population increases to $n+1$ and now the time interval until the next split is exponentially distributed with parameter $(n+1)\lambda$, etc. The study of this example from the point of view of pure birth processes was given in Section 1 of Chapter 4. The following alternative method is of independent interest.

When $1 - F(t) = e^{-\lambda t}$ Eq. (11.1) becomes

$$G(s, t)\, e^{\lambda t} = \lambda \int_0^t [G(s, t-\tau)]^2 e^{\lambda(t-\tau)}\, d\tau + s.$$

After executing the change of variables $u = t - \tau$ we get

$$G(s, t)e^{\lambda t} = \lambda \int_0^t [G(s, u)]^2 e^{\lambda u}\, du + s.$$

Now differentiate with respect to t. There results the equation

$$e^{\lambda t}G'(s, t) + \lambda e^{\lambda t}G(s, t) = \lambda[G(s, t)]^2 e^{\lambda t},$$

where

$$G'(s, t) = \frac{d}{dt} G(s, t).$$

Canceling the factor $e^{\lambda t}$ the differential equation reduces to a Bernoulli-type differential equation:

$$G'(s, t) = \lambda[G(s, t)]^2 - \lambda G(s, t). \tag{11.3}$$

To solve this differential equation we may simply separate variables:

$$\frac{dG(s, t)}{G(s, t)[G(s, t) - 1]} = \lambda\, dt.$$

Then the solution is

$$\frac{G(s,t)-1}{G(s,t)} = C(s)e^{\lambda t},$$

or explicitly

$$G(s,t) = \frac{1}{1-C(s)e^{\lambda t}}, \tag{11.4}$$

where $C(s)$ is a constant in t, but may be a function of s. To determine $C(s)$, let $t=0$ in (11.4). Since

$$p_k(0) = \begin{cases} 0 & \text{if } k \neq 1, \\ 1 & \text{if } k = 1, \end{cases} \qquad s \equiv G(s,0) = \frac{1}{1-C(s)},$$

it follows that

$$C(s) = \frac{s-1}{s}.$$

Thus, the solution of (11.3) and also the solution of (11.1) in the case of exponential lifetime is

$$G(s,t) = \frac{se^{-\lambda t}}{1-(1-e^{-\lambda t})s}. \tag{11.5}$$

To obtain explicit formulas for $p_k(t)$ we expand (11.5) in powers of s; i.e.,

$$G(s,t) = e^{-\lambda t}s \sum_{k=0}^{\infty} (1-e^{-\lambda t})^k s^k.$$

Visibly, we have

$$p_k(t) = e^{-\lambda t}(1-e^{-\lambda t})^{k-1} \qquad \text{for} \quad k = 1, 2, \ldots.$$

Although we cannot solve the integral equation (11.1) in the general case, we may obtain from it an equation for the mean function $m(t) = EN(t)$. Recall that

$$\left.\frac{dG(s,t)}{ds}\right|_{s=1} = \sum_{k=1}^{\infty} kp_k(t) = m(t).$$

Differentiating (11.1) with respect to s leads to

$$\frac{dG(s,t)}{ds} = 2\int_0^t G(s,t-\tau)\frac{dG(s,t-\tau)}{ds}f(\tau)\,d\tau + 1 - F(t).$$

Now, set $s=1$ on both sides, remembering that

$$G(1,t-\tau) = \textstyle\sum_{k=1}^{\infty} p_k(t-\tau) = 1.$$

Then

$$m(t) = 2\int_0^t m(t-\tau)f(\tau)\,d\tau + 1 - F(t).$$

This integral equation is an example of what is called a renewal equation. Its characteristic feature is the appearance of the unknown function under the integral sign in the convolution form. There is much classical theory available concerning the renewal equation which describes the asymptotic growth properties of the solution $m(t)$.

An obvious generalization of the above model is obtained by allowing an object at the time of its death to split into exactly r new objects of the same kind, where r is a fixed integer greater than 2. It is easy to see that the integral equation (11.1) will then be replaced by

$$G(s,t) = \int_0^t [G(s,t-\tau)]^r f(\tau)\,d\tau + [1 - F(t)]s.$$

A further generalization of this model is the case in which any object may split into a random number of new objects of the same kind at the time of its death; e.g., we may assume that an object produces l new objects at the time of its death with probability q_l, $l = 0, 1, 2, \ldots$. Let

$$h(s) = \sum_{l=0}^{\infty} q_l s^l$$

be the corresponding generating function. Then we may derive the integral equation in the following way. Assume that the first split occurs at time τ $(0 \leq \tau \leq t)$ and l new objects are created. This event has probability $f(\tau)\,d\tau\,q_l$. Then during the remaining $t - \tau$ units of time each of the l objects may produce any number of descendants such that the total number of objects at time t totals k. By the law of total probabilities, we have

$$p_k(t) = \int_0^t d\tau\, f(\tau) \sum_{l=0}^{\infty} q_l \sum_{k_1+k_2+\cdots+k_l=k} p_{k_1}(t-\tau)p_{k_2}(t-\tau)\cdot\cdots\cdot p_{k_l}(t-\tau)$$

for $k = 2, 3, \ldots$

and $p_1(t) = 1 - F(t)$ as before. Then the generating function

$$G(s,t) = [1 - F(t)]s + \sum_{k=2}^{\infty} s^k \int_0^t d\tau\, f(\tau) \sum_{l=0}^{\infty} q_l \sum_{k_1+\cdots+k_l=k} p_{k_1}(t-\tau)\cdots p_{k_l}(t-\tau)$$

$$= [1 - F(t)]s + \int_0^t d\tau\, f(\tau) \sum_{l=0}^{\infty} q_l \sum_{k=1}^{\infty} s^k \sum_{k_1+\cdots+k_l=k} p_{k_1}(t-\tau)\cdot\cdots\cdot p_{k_l}(t-\tau).$$

But the inner sum of

$$\sum_{k=0}^{\infty} s^k \sum_{k_1+\cdots+k_l=k} p_{k_1}(t-\tau)\cdot \cdots \cdot p_{k_l}(t-\tau)$$

is recognized as an l-fold convolution (the student should prove this) where each factor corresponds to

$$\sum_{k=0}^{\infty} s^k p_k(t-\tau) = G(s, t-\tau)$$

and the full sum is $[G(s, t-\tau)]^l$. Thus,

$$G(s, t) = [1-F(t)]s + \int_0^t d\tau\, f(\tau) \sum_{l=0}^{\infty} q_l[G(s, t-\tau)]^l.$$

But the summation inside the integral on the right-hand side gives the generating function $h(s)$ evaluated at $x = G(s, t-\tau)$. Finally, in this case the integral equation takes the form

$$G(s, t) = \int_0^t h(G(s, t-\tau))f(\tau)\, d\tau + [1 - F(t)]s.$$

Elementary Problems

1. Let X_n be a branching process where $X_0 \equiv 1$. For an arbitrary but fixed positive integer k define the sequence

$$Y_r = X_{rk}, \qquad r = 0, 1, 2, \ldots.$$

Show that $\{Y_r, r = 0, 1, 2, \ldots\}$ generates a branching process. Moreover, prove that if $\varphi(s)$ denotes the generating function of the number of direct descendants of a single individual in the X_n process and $\varphi_n(s)$ its nth iterate, then $\varphi_k(s)$ is the generating function in the Y_r process of the number of direct descendants of a single individual.

2. Let $f(s) = 1 - p(1-s)^\beta$, where p and β are constants and $0 < p < 1$, $0 < \beta < 1$. Prove that $f(s)$ is a probability generating function and its iterates are

$$f_n(s) = 1 - p^{1+\beta+\cdots+\beta^{n-1}}(1-s)^{\beta^n} \qquad \text{for} \quad n = 1, 2, \ldots.$$

3. Suppose $f(s)$ is a probability generating function and $h(s)$ is a function such that

$$g(s) = h^{-1}[f(h(s))]$$

is a probability generating function. Verify that

$$g_n(s) = h^{-1}[f_n(h(s))]$$

is a probability generating function, where f_n and g_n denote the functional iterates of f and g, respectively.

4. As an example of Elementary Problem 3, take

$$f(s) = \frac{s}{m-(m-1)s} \qquad (m>1)$$

and

$$h(s) = s^k \qquad (k \text{ a positive integer}).$$

Prove that $g(s) = h^{-1}f(h(s))$ is a generating function and establish that the nth iterate of g is

$$g_n(s) = \frac{s}{(m^n-(m^n-1)s^k)^{1/k}}.$$

5. Show that $E(\sum_{n=1}^{\infty} X_n) = m/(1-m)$ when $m = E(X_1) < 1$ in a branching process.

6. At time 0, a blood culture starts with one red cell. At the end of one minute, the red cell dies and is replaced by one of the following combinations with probabilities as indicated:

2 red cells	$\frac{1}{4}$
1 red, 1 white	$\frac{2}{3}$
2 white	$\frac{1}{12}$

Each red cell lives for one minute and gives birth to offspring in the same way as the parent cell. Each white cell lives for one minute and dies without reproducing. Assume the individual cells behave independently.

(a) At time $n+\frac{1}{2}$ minutes after the culture began, what is the probability that no white cells have yet appeared?

(b) What is the probability that the entire culture dies out eventually?

Solution: (a) $(\frac{1}{4})^{2^n-1}$; (b) $\frac{1}{3}$.

7. Let $f(s) = as^2 + bs + c$, where a, b, c are positive and $f(1) = 1$. Assume that the probability of extinction is d $(0 < d < 1)$. Prove that

$$d = \frac{c}{a}.$$

8. Suppose that in a branching process the number of offspring of an initial particle has a distribution whose generating function is $f(s)$. Each member of the first generation has a number of offspring whose distribution has generating function $g(s)$. The next generation has generating function f, the next g, and the functions continue to alternate in this way from generation to generation.

Arguing from basic principles (i.e., without using any general results from multi-type theory of Sections 5 and 6, determine extinction probability of the process, and the mean number of particles in the nth generation (n even, say). Would either of these quantities change if we started the process with the g function, and then continued to alternate?

9. Consider a discrete time branching process X_n with $X_0 = 1$. Establish the simple inequality

$$\Pr\{X_n > L \text{ for some } 0 \le n \le m | X_m = 0\} \le [\Pr\{X_m = 0\}]^L.$$

10. Consider a branching process with initial size N and probability generating function

$$\varphi(s) = q + ps, \qquad q, p > 0, \quad q + p = 1.$$

Determine the probability distribution of the time T when the population first becomes extinct.

Solution: $\Pr\{T = n\} = (1 - p^{n+1})^N - (1 - p^n)^N$.

11. Compute Var $X(t)$, where $X(t)$ is a continuous time branching process and $X(0) = 1$.

Solution:

$$\text{Var } X(t) = \begin{cases} \left[\dfrac{u''(1) - u'(1)}{u'(1)}\right] e^{u'(1)t}(e^{u'(1)t} - 1) & \text{if } u'(1) \neq 0, \\ u''(1)t & \text{if } u'(1) = 0. \end{cases}$$

12. A population consists of two types of individuals, males and females. We assume that all the females can produce offspring, according to a generating function $f(x)$, provided that the population contains at least one male. If the probability that an offspring is female is α, what is the p.g.f. for the number of females produced in the next generation given that at least one male is produced as well.

Solution: $\dfrac{f(\alpha s + (1 - \alpha)) - f(\alpha s)}{1 - f(\alpha)}$.

Problems

1. The following model has been introduced to study a urological process. Suppose bacteria grow according to a Yule process of parameter λ (see Section 1, Chapter 4). At each unit of time each bacterium present is eliminated with probability p. What is the probability generating function of the number of bacteria existing at time n?

Hint: This is the probability generating function of the nth iterate of a branching process.

Answer: $f_n(s)$ is the nth iterate of

$$f(s) = \frac{e^{-\lambda}(1 - e^{-\lambda})^{-1}}{1 - (1 - e^{-\lambda})(p + qs)}.$$

2. (a) A mature individual produces offspring according to the probability-generating function $f(s)$. Suppose we have a population of k immature individuals, each of which grows to maturity with probability p and then reproduces independently of the other individuals. Find the probability generating function of the number of (immature) individuals at the next generation.

(b) Find the probability generating function of the number of mature

individuals at the next generation, given that there are k mature individuals in the parent generation.

Answer: (a) $(1-p+pf(s))^k$; (b) $(f(1-p+ps))^k$.

3. Show that the distributions in (a) and (b) of Problem 2 have the same mean, but not necessarily the same variance.

4. Consider a discrete time branching process $\{X_n\}$ with probability generating function

$$\varphi(s)=\frac{1-(b+c)}{1-c}+\frac{bs}{1-cs}, \qquad 0<c<b+c<1,$$

where $(1-b-c)/c(1-c)>1$. Assume $X_0=1$. Determine the conditional limit distribution:

$$\lim_{n\to\infty} \Pr\{X_n=k|X_n>0\}.$$

Answer:

$$\left(1-\frac{1}{s_0}\right)\left(\frac{1}{s_0}\right)^{k-1}, \qquad s_0=\frac{1-b-c}{c(1-c)}.$$

5. In the previous problem suppose $1-b-c=c(1-c)$. Determine $\Pr\{X_n>0\}$.

Answer: $(1-c)/[1+(n-1)c]$.

6. Under the same conditions as in Problem 5 prove that $\Pr\{X_n\leq nx|X_n>0\}$ converges to an exponential distribution.

Hint: Compute the Laplace transform of X_n/n conditioned on the event $X_n>0$ and determine its limit as $n\to\infty$.

Answer: Exponential with parameter $(1-c)/c$.

7. Find the generating function $\varphi(t;s)$ of the continuous time branching process with infinitesimal generating function

$$u(s)=s^k-s \qquad (k\geq 2, \text{ integer}).$$

Hint: Solve

$$\frac{\partial\varphi(t;s)}{\partial t}=u(\varphi(t;s)) \qquad \text{with} \quad \varphi(0;s)=s.$$

Answer: $\varphi(t;s)=s[e^{(k-1)t}-(e^{(k-1)t}-1)s^{k-1}]^{-1/(k-1)}$.

8. Find the generating function $\varphi(t;s)$ of the continuous time branching process if the infinitesimal generating function is

$$u(s)=1-s-\sqrt{1-s}.$$

Answer: $\varphi(t;s)=1-[1-e^{-t/2}+e^{-t/2}\sqrt{1-s}]^2$.

9. Consider a multiple birth Yule process where each member in a population has a probability $\beta h+o(h)$ of giving birth to k new members and probability

$(1-\beta h+o(h))$ of no birth in an interval of time length h ($\beta>0$, k positive integer). Assume that there are N members present at time 0.

(a) Let $X(t)$ be the number of splits up to time t. Determine the growth behavior of $E(X(t))$.

(b) Let τ_n be the time of the nth split. Find the density function of τ_n.

Hint: (a) Note that

$$\Pr\{\tau_n \le t | \tau_{n-1}=\xi\} = \begin{cases} 1-\exp\{-[(n-1)k+N]\beta(t-\xi)\}, & \xi \le t \\ 0, & \xi > t \end{cases}$$

and obtain a recursion formula for the density function of τ_n in terms of the density function of τ_{n-1}.

Answer:

(a) $$\frac{E(X(t))}{e^{k\beta t}} \to \frac{N}{k} \quad \text{as} \quad t \to \infty.$$

(b) Let $f_n(t)$ be the probability density function of τ_n:

$$f_n(t) = \frac{N(N+k)\cdots[N+(n-1)k]}{(n-1)!k^{n-1}} \beta e^{-N\beta t}(1-e^{-k\beta t})^{n-1}.$$

10. Let X_n, $n \ge 0$, describe a branching process with associated probability generating function $\varphi(s)$.

Define Y_n as the total number of individuals in the first n generations, i.e.,

$$Y_n = X_0 + X_1 + \cdots + X_n, \qquad n=0,1,2,\ldots, \qquad X_0=1,$$

Let $F_n(s)$ be the probability generating function of Y_n. Establish the functional relation

$$F_{n+1}(s) = s\varphi(F_n(s)), \qquad \text{for} \quad n=0,1,2,\ldots.$$

11. Let $\varphi(s)$ be the generating function of the number of progeny of a single individual in a branching process that starts with one individual at time zero, and let $\varphi_n(s)$ denote its nth iterate.

Suppose in addition to the ordinary branching process there also exists some immigration into the population during a single generation described by the probability generating function $h(s)$. Consider the branching process with immigration whose transition probability matrix is defined by

$$\sum_{j=0}^{\infty} P_{ij}s^j = [\varphi(s)]^i \cdot h(s).$$

Prove that the n-step transition probability matrix is determined by the relation

$$\sum_{j=0}^{\infty} P_{ij}^n s^j = [\varphi_n(s)]^i h(\varphi_{n-1}(s))h(\varphi_{n-2}(s))\cdot\cdots\cdot h(\varphi(s))h(s).$$

12. In the branching process with immigration (Problem 11) assume that $\varphi'(1)=m<1$. Prove that the associated Markov chain has a stationary probability distribution with probability generating function $\pi(s) = \sum_{r=0}^{\infty} \pi_r s^r$

that satisfies the functional equation

$$\pi(\varphi(s))h(s) = \pi(s).$$

13. Under the set-up of Problem 12 for the specification $\varphi(s) = q + ps$ $(0 < p < 1, q + p = 1)$ and $h(s) = e^{s-1}$ determine the stationary probability distribution.

14. Consider the simple birth and death process (linear growth without immigration), i.e., $\lambda_n = \lambda n$ and $\mu_n = \mu n$ with $\lambda > 0$, $\mu > 0$, and $\mu > \lambda$. Let $Z(t)$ be the population size at time t. By appropriate identifications, show that the busy period of an infinite server queueing process with the interarrival distribution $1 - e^{-\lambda t}$ and the service time distribution $1 - e^{-\mu t}$ has the same distribution as that of $\int_0^\infty Z(t)\, dt$ under the initial condition $Z(0) = 1$.

15. Let X_n be a discrete branching process with associated probability generating function $\varphi(s)$ and let $\varphi_n(s) = \sum_{k=0}^{\infty} \Pr\{X_n = k\}\, s^k$. Assume that $\varphi'(1) > 1$. Let

$\tilde{X}_n$ denote the number of all the particles in the nth generation which have an infinite line of descent.

Show that the probability generating function for $\tilde{X}_n$ is

$$\sum_{k=0}^{\infty} \Pr\{\tilde{X}_n = k \,|\, \tilde{X}_0 = X_0 = 1\}\, s^k = \frac{\varphi_n(s(1-q)+q) - q}{1-q}$$

where q is the probability of extinction.

Hint: Note that for $k \geq 1$

$$\Pr\{\tilde{X}_n = k \,|\, \tilde{X}_0 = 1, X_0 = 1\} = \frac{\sum_{l=k}^{\infty} \Pr\{\tilde{X}_n = k, X_n = l \,|\, X_0 = 1\}}{\Pr\{\tilde{X}_0 = 1 \,|\, X_0 = 1\}}.$$

16. The purpose of this next problem is to determine the effects that different forms of mortality have on the stability of a population. We define *stability* as the probability of indefinite survivorship = 1 − probability of eventual extinction.

In the absence of the additional mortality we'll consider momentarily, the offspring X of a single individual has the probability distribution

$$\Pr\{X = k\} = p_k, \quad k = 0, 1, \ldots.$$

Suppose that the mean of the distribution is m and that all offspring in the population are independent and identically distributed.

We consider 3 types of mortality. In each case, the probability of an individual surviving is p, but the form the survivorship takes differs among the cases. Assume

$$mp > 1.$$

(a) *Mortality on Individuals:* Independent of what happens to others, each individual survives with probability p. That is, given an actual litter size or

number of offspring of X, the effective litter size has a binomial distribution with parameters (X, p). This type of mortality might reflect predation on adults.

(b) *Mortality on Litters:* Independent of what happens to other litters, each litter survives with probability p and is wiped out with probability $q = 1 - p$. That is, given an actual litter size of X, the effective litter size is X with probability p and 0 with probability q. This type of mortality might reflect predation on juveniles, or on nests and eggs in the case of birds.

(c) *Mortality on Generations:* An entire generation survives with probability p and is wiped out with probability q. This type of mortality might represent environmental catastrophies such as forest fire, flood, etc.

Give the equations for determining 1 — Stability = Pr {Eventual Extinction} in each of these cases.

Which population is the most stable? Which is least stable? Can you prove this?

NOTES

The source of inspiration for this chapter is the treatise on branching processes by T. Harris [1], which also contains a comprehensive bibliography of the subject and its applications.

REFERENCE

1. T. Harris, "The Theory of Branching Processes." Springer-Verlag, Berlin, 1963.

Chapter 9

STATIONARY PROCESSES

A stationary process is a stochastic process whose probabilistic laws remain unchanged through shifts in time (sometimes in space). The concept captures the very natural notion of a physical system that lacks an inherent time (space) origin. It is an appropriate assumption for a variety of processes in communication theory, astronomy, biology, ecology, and economics.

The stationary property leads to a number of important conclusions in a rich theory. In this chapter we focus on the prediction problem, the ergodic behavior, the spectral representation of a stationary process, stationary point processes, and the level-crossing problem. Sections 1 and 2 are prerequisite to the later sections. However, the section pairs 3 and 4, on prediction; 5 and 6, on ergodic theory; 7 and 8, on spectral analysis; and 9 and 10 on point processes and the level-crossing problem may be read in any order one desires.

1: Definitions and Examples

Let T be an abstract index set having the property that the sum of any two points in T is also in T. Often T will be the set $\{0, 1, \ldots\}$ of nonnegative integers, but it just as well could be the positive half or whole real line, the plane, finite-dimensional space, the surface of a sphere, or perhaps even an infinite-dimensional space.

Definition 1.1. *A stationary process is a stochastic process* $\{X(t),\ t \in T\}$ *with the property that for any positive integer* k *and any points* $t_1, \ldots, t_k$ *and* h *in* T, *the joint distribution of*

$$\{X(t_1), \ldots, X(t_k)\}$$

is the same as the joint distribution of

$$\{X(t_1 + h), \ldots, X(t_k + h)\}.$$

Here are some short examples.

(a) Electrical pulses in communication theory are often postulated to describe a stationary process. Of course, in any physical system there is a transient period at the beginning of a signal. Since typi-

cally this has a short duration compared to the signal length, a stationary model may be appropriate. In electrical communication theory, often both the electrical potential and the current are often represented as complex variables. Here we may encounter complex-valued stationary processes.

(b) The spatial and/or planar distributions of stars or galaxies, plants, and animals, are often stationary. Here T might be Euclidean space, the surface of a sphere, or the plane.

A stationary distribution may be postulated for the height of a wave and T taken to be a set of longitudes and latitudes, again two dimensional.

(c) Economic time series, such as unemployment, gross national product, national income, etc., are often assumed to correspond to a stationary process, at least after some correction for long-term growth has been made.

As these examples show, stationary processes appear in an abundance of shapes and sizes. To treat the most general situation would counter our purpose of providing an introduction. Having alerted the reader to the vast scope of possibilities, henceforth we concentrate mostly on the simplest case of a real-valued process for which $T = \{0, 1, 2, \ldots\}$.

Let $\{X(t),\ t \in T\}$ be a stationary process. If the mean $m(t) = E[X(t)]$ exists, it follows that this quantity must be a constant, $m(t) = m$ for all t. Similarly, if the second moment $E[X(t)^2]$ is finite, then the variance $\sigma^2 = E[(X(t) - m)^2]$ is a constant, independent of time. Let t and s be time points, and suppose, without loss in generality, that $t > s$. Using the stationary property, we compute the covariance

$$E[(X(t) - m)(X(s) - m)] = E[(X(t - s) - m)(X(0) - m)],$$

such that the right-hand side depends only on the time difference $t - s$. If we define the *covariance function*,

$$R(h) = E[(X(h) - m)(X(0) - m)],$$

then

$$E[(X(t) - m)(X(s) - m)] = R(|t - s|).$$

Of course, $\sigma^2 = R(0)$. Sometimes it is convenient to standardize the covariance producing what is called the *correlation function* or *autocorrelation function*, defined by

$$\rho(v) = \frac{1}{\sigma^2} R(v) = R(v)/R(0).$$

Then $\rho(0) = 1$, and it can be shown (using Schwarz' inequality) that $-1 \leq \rho(v) \leq 1$ for all v.

The concept of stationarity introduced in Definition 1.1 involves all finite-dimensional distributions of the process. For many purposes it is desirable to have available a weaker concept, involving only the first two moments.

Definition 1.2. *A covariance stationary process is a stochastic process* $\{X(t), t \in T\}$ *having finite second moments,* $E[X(t)^2] < \infty$, *a constant mean* $m = E[X(t)]$, *and a covariance* $E[(X(t) - m)(X(s) - m)]$ *that depends only on the time difference* $|t - s|$.

Other terms used in the literature synonymously with covariance stationary are *weakly stationary* or *wide-sense stationary*, and what we have called a stationary process is often termed *strictly stationary* to emphasize the distinction.

A stationary process that has finite second moments is covariance stationary (but, of course, a stationary process may have no finite moments whatsoever). It is quite possible that a covariance stationary process will not be stationary, but there is an important exception to this general rule. A stochastic process $\{X(t), t \in T\}$ for which, for every k and every finite set $\{t_1, \ldots, t_k\}$ of time points, the random vector

$$(X(t_1), \ldots, X(t_k))$$

has a multivariate normal distribution (Chapter 1) is called a *Gaussian process*. Since the multivariate normal distribution is determined by its first two moments, the mean value vector and the covariance matrix, a Gaussian process which is covariance stationary will be strictly stationary.

Examples

Several of the examples that follow will be developed further in the sequel.

A. *Two Contrasting Stationary Processes*

(i) A sequence $Y_0, Y_1, \ldots$ of independent and identically distributed random variables is a stationary process. If the common distribution of $Y_0, Y_1, \ldots$ has a finite variance σ^2 then the process is covariance stationary, and the covariance function is

$$R(v) = \begin{cases} \sigma^2, & \text{for } v = 0, \\ 0, & \text{for } v \neq 0. \end{cases}$$

(ii) To consider a quite different stationary process, let Z be a single random variable with known distribution and set $Z_0 = Z_1 = Z_2 = \cdots = Z$. The process $\{Z_n\}$ is easily seen to be stationary. If the random variable Z has a finite variance σ^2, then the process is covariance stationary, and the covariance function is

$$R(v) = \sigma^2, \qquad \text{for all} \quad v.$$

In many ways $\{Y_n\}$ and $\{Z_n\}$ are extremes and may be used to exemplify contrasting behavior of stationary processes. For example, assuming that the common distribution is known, observing $Y_0, Y_1, \ldots, Y_n$ provides no information that could be used to predict Y_{n+1}, while observing only Z_0 enables $Z_1, Z_2, \ldots$ to be predicted exactly. Here is a second way that the processes are opposites. Suppose the Y_n process has a finite mean value function m. Then by the law of large numbers, the sample averages

$$\frac{1}{n}(Y_0 + \cdots + Y_{n-1})$$

converge to the constant $m = E[Y_0]$. No such convergence takes place in the $\{Z_n\}$ process. Indeed

$$\frac{1}{n}\{Z_0 + \cdots + Z_{n-1}\} = Z_0 = Z,$$

and there is just as much "randomness" in the nth sample average as there is in the first observation.

The behavior in which sample averages formed from a process converge to some underlying parameter of the process is termed *ergodic*. To make inferences about the underlying laws governing an ergodic process, one need not observe separate independent replications of entire processes or sample paths. Instead, one need only observe a single realization of the process, but over a sufficiently long span of time. Thus, it is an important practical problem to determine conditions that lead to a stationary process being ergodic.

Perhaps surprisingly, these two examples of opposite behavior have related causes. For covariance stationary processes, the crux of the matter in both situations is whether or not the covariance function $R(|t - s|)$ converges to zero as the time difference $|t - s|$ becomes large, and if it does so vanish, the rate at which this convergence takes place has relevance. For the $\{Y_n\}$ process, the convergence is very fast indeed, since the covariance is exactly zero for lags or time differences of one or more, while for the $\{Z_n\}$ process, the correlation function is one at all time differences.

The theory of stationary processes has as a prime goal the clarification of ergodic behavior and the prediction problem for processes falling in the vast range between the two extreme examples just exhibited. The general problem of prediction for stationary processes is studied in Sections 3 and 4 of this chapter, and the convergence of sample averages is elaborated in Sections 5 and 6.

B. *Trigonometric Polynomials*

Some interesting examples of stationary processes can be obtained by considering certain trigonometric expressions having random amplitudes. Let A and B be identically distributed random variables having a mean of zero and variance σ^2. We suppose A and B are uncorrelated, i.e., $E[AB] = 0$. Fix a frequency $\omega \in [0, \pi]$ and for $n = 0, \pm 1, \pm 2, \ldots$ define

$$X_n = A \cos(\omega n) + B \sin(\omega n).$$

Then $E[X_n] = 0$ for all n, and, using the trigonometric identity

$$\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta,$$

and the fact that $E[AB] = 0$, we compute the covariance

$$\begin{aligned} E[X_n X_{n+v}] &= E[\{A \cos \omega n + B \sin \omega n\}\{A \cos \omega(n+v) + B \sin \omega(n+v)\}] \\ &= E[A^2 \cos \omega n \cos \omega(n+v) + B^2 \sin \omega n \sin \omega(n+v)] \\ &= \sigma^2 \cos \omega v. \end{aligned}$$

Since the covariance between X_n and X_{n+v} plainly depends only on the time difference v, we conclude that the process is covariance stationary. If A and B have a normal distribution with mean zero and variance σ^2, then the process is Gaussian and thus strictly stationary.

For the particular frequency $\omega = 0$, we have $\cos n\omega = 1$ and $\sin n\omega = 0$, so that $X_n = A$ for all n. Thus the $\{Z_n\}$ process in the previous example falls within this framework.

More generally, let $A_0, A_1, \ldots, A_m$ and $B_0, B_1, \ldots, B_m$ be uncorrelated random variables having zero means. Assume that A_i and B_i have a common variance σ_i^2, and let $\sigma^2 = \sigma_0^2 + \cdots + \sigma_m^2$. Take $\omega_0, \omega_1, \ldots, \omega_m$ as distinct frequencies in $[0, \pi]$, and for $n = 0, \pm 1, \pm 2, \ldots$ set

$$X_n = \sum_{k=0}^{m} \{A_k \cos n\omega_k + B_k \sin n\omega_k\}.$$

Since the coefficients $\{A_k\}$ and $\{B_k\}$ are uncorrelated with zero means, we have $E[A_i B_j] = 0$ and $E[A_i A_k] = E[B_i B_k] = 0$ for $k \neq i$. We compute the covariance

$$\begin{aligned} E[X_n X_{n+v}] &= E\left[\left(\sum_{k=0}^{m} \{A_k \cos n\omega_k + B_k \sin n\omega_k\}\right)\right. \\ &\quad \left.\cdot\left(\sum_{j=0}^{m} \{A_j \cos(n+v)\omega_j + B_j \sin(n+v)\omega_j\}\right)\right] \\ &= \sum_{k=0}^{m} E[A_k^2 \cos n\omega_k \cos(n+v)\omega_k + B_k^2 \sin n\omega_k \sin(n+v)\omega_k] \\ &= \sum_{k=0}^{m} \sigma_k^2 \cos v\omega_k. \end{aligned}$$

Again, the process is covariance stationary.

To go on, it is helpful to let $p_k = \sigma_k^2/\sigma^2$ and write the covariance function as

$$R(v) = \sigma^2 \sum_{k=0}^{m} p_k \cos v\omega_k. \tag{1.1}$$

Then p_k represents the contribution of frequency ω_k in the covariance. Observe that $\{p_k\}$ is a discrete probability distribution, that is, $p_k \geq 0$ and $\sum p_k = 1$, which suggests the possibility of generalizing (1.1) to a continuum of frequencies, something of the form

$$R(v) = \sigma^2 \int_0^{\pi} \cos(v\omega)\, dF(\omega), \tag{1.2}$$

where $F(\omega)$ is a cumulative distribution function of a random variable having possible values in $[0, \pi]$. In Section 7, we shall see that such a generalization is indeed possible and that the most general covariance stationary process has a representation of this form. In the special case in which F corresponds to a uniform distribution on $[0, \pi]$, meaning all frequencies are equally represented, we calculate

$$\begin{aligned} R(v) &= \sigma^2 \frac{1}{\pi} \int_0^{\pi} \cos(v\omega)\, d\omega \\ &= \begin{cases} \sigma^2, & \text{if } v = 0, \\ 0, & \text{if } v \neq 0. \end{cases} \end{aligned}$$

This is the covariance function of the independent and identically distributed sequence $\{Y_n\}$ in the previous example. Again $\{Y_n\}$ and $\{Z_n\}$

are opposites in some sense, $\{Z_n\}$ corresponding to a single frequency $\omega = 0$, and $\{Y_n\}$ corresponding to all frequencies in $[0, \pi]$ given equal weight.

C. *Moving Average Processes*

Let $\{\xi_n : n = 0, \pm 1, \pm 2, \ldots\}$ be uncorrelated random variables having a common mean μ and variance σ^2. Let $a_1, a_2, \ldots, a_m$ be arbitrary real numbers and consider the process defined by

$$X_n = a_1\xi_n + a_2\xi_{n-1} + \cdots + a_m\xi_{n-m+1}.$$

We have

$$E[X_n] = \mu(a_1 + \cdots + a_m),$$

and

$$\mathrm{Var}[X_n] = \sigma^2(a_1^2 + \cdots + a_m^2),$$

for the mean and variance, respectively. Let $\hat{\xi}_k = \xi_k - \mu$. For the covariance, we have

$$\begin{aligned} &E\left[\left(X_n - \mu\sum_{i=1}^{m} a_i\right)\left(X_{n+v} - \mu\sum_{i=1}^{m} a_i\right)\right] \\ &\quad = E\left[\left(\sum_{i=1}^{m} a_i\hat{\xi}_{n-i+1}\right)\left(\sum_{j=1}^{m} a_j\hat{\xi}_{n+v-j+1}\right)\right] \\ &\quad = \begin{cases} E[a_m a_{m-v}\hat{\xi}^2_{n+v-m+1} + a_{m-1}a_{m-v-1}\hat{\xi}^2_{n+v-m+2} + \cdots \\ \qquad + a_{v+1}a_1\hat{\xi}^2_n], & \text{if } v \le m-1, \\ 0, & \text{if } v \ge m. \end{cases} \\ &\quad = \begin{cases} \sigma^2(a_m a_{m-v} + \cdots + a_{v+1}a_1), & \text{if } v \le m-1, \\ 0, & \text{if } v \ge m. \end{cases} \end{aligned}$$

Since the covariance between X_n and X_{n+v} depends only on the lag v, and not on n, the process is covariance stationary.

A common case is the "moving average" with a standardized variance in which $a_k = 1/\sqrt{m}$ for $k = 1, \ldots, m$. The covariance function becomes

$$R(v) = \begin{cases} \sigma^2\left(1 - \dfrac{v}{m}\right), & v \le m-1, \\ 0, & \text{for } v \ge m. \end{cases}$$

The case $m = 1$ corresponds to the uncorrelated random variables in the $\{Y_n\}$ process of Example A, and, at the other extreme, roughly speaking, $m = \infty$ corresponds to the $\{Z_n\}$ process.

D. *A Stationary Process on the Circle*

Let U, V be independent normally distributed random variables having zero mean and unit variance. Let T be the circumference of the unit circle, represented by $T = [0, 2\pi]$, and for $t \in T$ define the bivariate process $X(t) = (Y(t), Z(t))$, where

$$Y(t) = U \sin t + V \cos t,$$
$$Z(t) = -U \cos t + V \sin t.$$

Then $(X(t), t \in T)$ is a stationary process. Clearly, $E[Y(t)] = E[Z(t)] = 0$, and since $\sin^2 t + \cos^2 t = 1$ for all t, $E[Y(t)^2] = E[Z(t)^2] = 1$. Since $Y(t)$ and $Z(t)$ have a joint normal distribution, to complete the specification of their distribution, we need only compute their covariance. But, easily

$$E[Y(t)Z(t)] = 0.$$

Thus the distribution of $X(t)$ is the same as the distribution of $X(t + \theta)$ for any θ. Since one can verify the same property for any vector $(X(t_1), \ldots, X(t_k))$, the process is stationary.

E. *Stationary Markov Chains*

In Chapter 3 we showed that, under quite general conditions a Markov chain $\{X_n\}$ would evolve towards an equilibrium regime of statistical fluctuations in which the importance of the period n and initial state X_0 would have faded into the past. To be more precise, we gave conditions under which, as n became large, the distribution $\Pr\{X_n = j | X_0 = i\}$ would approach a constant π_j not depending on i. Let us suppose we are observing such a system, but one which began its evolution indefinitely far in the past and is now evolving in its equilibrium regime. Then for any $n = 0, \pm 1, \pm 2, \ldots$, the probability distribution of X_n does not depend on n. Indeed, $\Pr\{X_n = j\} = \pi_j$, so that this probability, or marginal distribution for X_n, is *stationary* in the sense that it does not depend on n. Similarly, the joint distribution of (X_n, X_{n+1}) does not depend on n, but is given by

$$\Pr\{X_n = i, X_{n+1} = j\} = \Pr\{X_n = i\} \Pr\{X_{n+1} = j | X_n = i\} = \pi_i P_{ij},$$

where $\mathrm{P} = \|P_{ij}\|$ is the transition matrix for the Markov chain. Quite obviously we may continue and state that for any fixed $k = 0, 1, \ldots$ the joint distribution of $(X_n, X_{n+1}, \ldots, X_{n+k})$ does not depend on n.

Similarly, let $\{X_n, n = 0, 1, \ldots\}$ be a Markov chain for which the initial state X_0 is chosen according to the stationary distribution π_j. The same reasoning shows that $\{X_n\}$ is a stationary process.

2: *Mean Square Distance*

Since "covariance stationarity" is a property defined using only the first two moments of a stochastic process, it is desirable to have a measure of dissimilarity or "distance" between random variables Y, Z, that also is defined in terms of only the first two moments. A natural choice for such a measure is the mean of the squared difference, $E[(Y-Z)^2]$, or what is even better, since it resumes the original units of measurement, the square root of this mean square difference. To enhance the suggestion of distance, we introduce the notation

$$\|Z\| = \sqrt{E[Z^2]}, \qquad \text{and} \qquad \|Y-Z\| = \sqrt{E[(Y-Z)^2]}.$$

Thus, for example, we would measure the ability of a predictor $\hat{X}_{t+k}$ to predict a random variable X_{t+k} by the root mean square difference $\|\hat{X}_{t+k} - X_{t+k}\|$. Again, using this measure of distance we may introduce a notion of convergence for sequences of random variables. Accordingly, a sequence $\{X_n\}$ of random variables will be said to converge to a random variable X if the mean square difference $E[(X_n - X)^2]$ tends to zero as n increases indefinitely. A formal definition follows:

Definition 2.1. *Let X, X_1, X_2, ... be random variables. We say $\{X_n\}$ converges in mean square to X, written $X_n \to X$ (ms) or $\lim_{n\to\infty} X_n = X$ (ms) if*

(i) $E[X_n^2] < \infty$ for all n, and
(ii) $\lim_{n\to\infty} E[(X_n - X)^2] = 0$ (or equivalently $\lim_{n\to\infty} \|X_n - X\| = 0$).

Here are some elementary properties of mean square distance and mean square convergence. In this list, Y, Z is an arbitrary pair of random variables having finite second moments, and y, z are arbitrary real numbers.

Schwarz' Inequality

Observe

$$2|yz| \leq y^2 + z^2. \tag{2.1}$$

Since this inequality is true for arbitrary y, z, it must hold when these values are chosen randomly, i.e.,

$$2|YZ| \leq |Y|^2 + |Z|^2.$$

By taking the expectation of both sides, we conclude

$$2E[|YZ|] \leq E[|Y|^2] + E[|Z|^2].$$

In particular, $E[|YZ|] < \infty$. If instead we substitute

$$Y/\sqrt{E[|Y|^2]} = Y/\|Y\|, \qquad \text{for} \quad y,$$

and

$$Z/\sqrt{E[|Z|^2]} = Z/\|Z\|, \qquad \text{for} \quad z,$$

in (2.1) and take expectations, we conclude

$$2E\left\{\frac{|YZ|}{\|Y\| \cdot \|Z\|}\right\} \leq E\left\{\frac{Y^2}{\|Y\|^2}\right\} + E\left\{\frac{Z^2}{\|Z\|^2}\right\} = 2$$

Hence

$$E[|YZ|] \leq \sqrt{E[Y^2]E[Z^2]} = \|Y\| \, \|Z\|.$$

This is known as *Schwarz' inequality*.

The Parallelogram Law

We compute

$$E[(Y+Z)^2] = E[Y^2 + 2YZ + Z^2] = E[Y^2] + 2E[YZ] + E[Z^2],$$

and

$$E[(Y-Z)^2] = E[Y^2 - 2YZ + Z^2] = E[Y^2] - 2E[YZ] + E[Z^2],$$

which added together give the parallelogram law

$$E[(Y+Z)^2] + E[(Y-Z)^2] = 2E[Y^2] + 2E[Z^2],$$

or

$$\|Y+Z\|^2 + \|Y-Z\|^2 = 2\{\|Y\|^2 + \|Z\|^2\}. \tag{2.2}$$

The name results from the similarity of this equation with the result in geometry, where Y, Z are vectors, that the sum of squares of the two diagonals in a parallelogram is equal to the sum of squares of the four sides.

The Triangle Inequality

We compute

$$E[(Y+Z)^2] = E[Y^2] + 2E[YZ] + E[Z^2].$$

From Schwarz' inequality,

$$E[YZ] \leq E[|YZ|] \leq \sqrt{E[Y^2]E[Z^2]}.$$

Hence

$$\|Y+Z\|^2 \le E[Y^2]+2\sqrt{E[Y^2]}\sqrt{E[Z^2]}+E[Z^2]$$
$$=(\|Y\|+\|Z\|)^2,$$

or

$$\|Y+Z\| \le \|Y\|+\|Z\|. \tag{2.3}$$

The name results from the similarity to the geometric inequality that states that the sum of lengths of two sides of a triangle always exceeds the length of the third side. Observe the parallel between $\|Z\|$ and the usual notion of length of a vector in space. There is an alternative form that is sometimes more useful:

$$|(\|Y\|-\|X\|)| \le \|Y-X\|. \tag{2.4}$$

This is obtained from (2.3) in the same way that $|(|a|-|b|)| \le |a-b|$ is obtained from $|a+b| \le |a|+|b|$ for real a, b.

Uniqueness of ms Limit

If we set $Y=X_n-X'$ and $Z=X-X_n$ in the triangle inequality (2.3), we get

$$\|X-X'\| \le \|X_n-X'\|+\|X_n-X\|.$$

Thus, if $X_n \to X$(ms) and $X_n \to X'$(ms), then $E[(X-X')^2]=\|X-X'\|^2 = 0$. Applying Chebyshev's inequality, we have

$$\Pr\{|X-X'|>\varepsilon\} \le \frac{1}{\varepsilon^2}E[(X-X')^2], \qquad \varepsilon>0,$$

yielding $\Pr\{|X-X'|>\varepsilon\}=0$ for all positive ε, and thus,

$$\Pr\{X=X'\}=1. \tag{2.5}$$

Hence the limit of a sequence of random variables converging in mean square is unique in the sense of (2.5).

Convergence in Probability

Again using Chebyshev's inequality

$$\Pr\{|X_n-X|>\varepsilon\} \le \frac{1}{\varepsilon^2}E[(X_n-X)^2], \qquad \varepsilon>0,$$

we see that $X_n \to X$(ms) implies

$$\lim_{n\to\infty}\Pr\{|X_n-X|>\varepsilon\}=0,$$

for any positive ε. Hence convergence in mean square implies convergence in probability. Warning! The converse is *not* true.

Convergence of Second Moments

Let us apply the triangle inequality (2.4) with $Y = X_n$. We have

$$|\,\|X\| - \|X_n\|\,| \leq \|X - X_n\| \to 0.$$

Hence

$$\lim_{n\to\infty} \|X_n\| = \|X\|. \tag{2.6}$$

The Cauchy Criterion for Convergence

If we set $Y = X_n - X$ and $Z = X - X_m$ in (2.2), we get

$$E[(X_n - X_m)^2] \leq 2E[(X_n - X)^2] + 2E[(X_m - X)^2].$$

Hence, if $X_n \to X$(ms) then

$$\lim_{m,\,n\to\infty} E[(X_n - X_m)^2] = 0, \qquad \text{or} \qquad \lim_{m,\,n\to\infty} \|X_n - X_m\| = 0.$$

Conversely, it can be shown that if $\{X_n\}$ is a sequence of random variables for which

$$\lim_{m,\,n\to\infty} E[(X_n - X_m)^2] = 0,$$

then there exists a random variable X for which

$$X_n \to X(\text{ms}).$$

This result parallels the Cauchy criterion for convergence of sequences of real numbers.

Convergence of Means

We have

$$E[|\,Y\,|] = E[|\,Y \cdot 1|] \leq \sqrt{E[Y^2] \cdot 1} = \|Y\|,$$

by using Schwarz' inequality with $Z \equiv 1$. Thus $E[|\,Y\,|] < \infty$ if $E[Y^2] < \infty$. In a similar manner, setting $Y = X_n - X$, we compute

$$|E[X_n] - E[X]| \leq E[|X_n - X|] \leq \|X_n - X\|.$$

Thus $X_n \to X$(ms) as $n \to \infty$ implies

$$\lim_{n\to\infty} E[X_n] = E[X].$$

Convergence of Covariances

Let $\{X_n\}$ be a sequence of random variables converging in mean square to a random variable X. Let Y be a random variable having a finite second

moment. We want to show $\lim_{n\to\infty} E[YX_n] = E[YX]$, or, equivalently, $\lim_{n\to\infty} |E[YX_n] - E[YX]| = 0$. We use Schwarz' inequality:

$$|E[YX_n] - E[YX]| = |E[Y(X_n - X)]| \leq \|Y\| \, \|X_n - X\|.$$

Since $E[Y^2] < \infty$ and $\|X_n - X\| \to 0$, the proof is complete.

AUTOREGRESSIVE AND MOVING AVERAGE PROCESSES

Let $\{X_n;\ n = 0, \pm 1, \pm 2, \ldots\}$ be a covariance stationary process. Suppose that for some real number λ, satisfying $|\lambda| < 1$, that the random variables defined by

$$\xi_n = X_n - \lambda X_{n-1}$$

are uncorrelated with zero means and a common variance σ^2. Such a process is called an *autoregressive process of order one.* We may write

$$\begin{aligned} X_n &= \lambda X_{n-1} + \xi_n \\ &= \lambda\{\lambda X_{n-2} + \xi_{n-1}\} + \xi_n \\ &= \lambda^2 X_{n-2} + \lambda\xi_{n-1} + \xi_n \end{aligned}$$

and inductively

$$= \lambda^k X_{n-k} + \sum_{j=0}^{k-1} \lambda^j \xi_{n-j}. \tag{2.7}$$

We rearrange terms and compute the mean square difference between X_n and $\sum_{j=0}^{k-1} \lambda^j \xi_{n-j}$. This yields

$$\begin{aligned} E\left[\left(X_n - \sum_{j=0}^{k-1} \lambda^j \xi_{n-j}\right)^2\right] &= E[(\lambda^k X_{n-k})^2] \\ &= \lambda^{2k} E[X_{n-k}^2]. \end{aligned}$$

Since the process is stationary, $E[X_{n-k}^2]$ is a constant, independent of n and k, and since $|\lambda| < 1$, the right-hand side decreases to zero at a geometric rate. Thus

$$\begin{aligned} X_n &= \lim_{k\to\infty} \sum_{j=0}^{k-1} \lambda^j \xi_{n-j} \text{ (ms)} \\ &= \sum_{j=0}^{\infty} \lambda^j \xi_{n-j}. \end{aligned} \tag{2.8}$$

It is to be remembered that $\sum_{j=0}^{\infty}$ signifies the limit *in mean square* of the sequence of partial sums $\sum_{j=0}^{k-1}$. Equation (2.8) provides a representation of the original process as a *moving average* (Example C, earlier in this section).

Since mean square convergence implies convergence of the means and second moments, we have

$$E[X_n] = \lim_{k\to\infty} E\left[\sum_{j=0}^{k-1} \lambda^j \xi_{n-j}\right] = 0,$$

and

$$\begin{aligned} E[X_n^2] &= \lim_{k\to\infty} E\left[\left(\sum_{j=0}^{k-1} \lambda^j \xi_{n-j}\right)^2\right] \\ &= \lim_{k\to\infty} \left\{E\left[\sum_{j=0}^{k-1} \lambda^{2j} \xi_{n-j}^2\right] + E\left[\sum_{i\neq j} \lambda^{i+j} \xi_{n-i}\, \xi_{n-j}\right]\right\}. \end{aligned}$$

The expectation of the second term vanishes since the $\{\xi_m\}$ sequence is uncorrelated. Since $E[\xi_{n-j}^2] = \sigma^2$ and $|\lambda| < 1$, we get

$$\begin{aligned} E[X_n^2] &= \lim_{k\to\infty} \sum_{j=0}^{k-1} \lambda^{2j}\sigma^2 \\ &= \sigma^2/(1 - |\lambda|^2). \end{aligned}$$

Let us compute the covariance between X_n and X_{n+k}. We have from (2.7)

$$X_{n+k} = \lambda^k X_n + \sum_{j=0}^{k-1} \lambda^j \xi_{n+k-j},$$

so that

$$E[X_n X_{n+k}] = \lambda^k E[X_n^2] + E\left[X_n \cdot \sum_{j=0}^{k-1} \lambda^j \xi_{n+k-j}\right].$$

The first term on the right is $\lambda^k\sigma^2$. Using the convergence of covariances, the second term is evaluated routinely, viz.,

$$\begin{aligned} E\left[X_n \sum_{j=0}^{k-1} \lambda^j \xi_{n+k-j}\right] &= \lim_{l\to\infty} E\left[\sum_{m=0}^{l-1} \lambda^m \xi_{n-m} \cdot \sum_{j=0}^{k-1} \lambda^j \xi_{n+k-j}\right] \\ &= \lim_{l\to\infty} \sum_{m=0}^{l-1} \sum_{j=0}^{k-1} \lambda^{m+j} E[\xi_{n-m}\, \xi_{n+k-j}] \\ &= 0, \end{aligned}$$

where the fact that $\{\xi_m\}$ are uncorrelated is heavily exploited. Thus

$$E[X_n X_{n+k}] = \sigma^2 \lambda^k, \qquad k = 0, 1, 2, \ldots.$$

To move on, writing the process in the form

$$X_n = \lambda X_{n-1} + \xi_n$$

suggests the natural generalization to covariance stationary processes that have the form

$$X_n = \lambda_1 X_{n-1} + \lambda_2 X_{n-2} + \cdots + \lambda_p X_{n-p} + \xi_n, \tag{2.9}$$

where $\{\xi_n\}$ is a sequence of zero mean uncorrelated random variables having a common variance. Such a process is called a pth *order autoregressive process.* For such a process, linear regression theory suggests that a natural predictor for X_n given the past $X_{n-1}, X_{n-2}, \ldots$ would be given by the formula

$$\hat{X}_n = \lambda_1 X_{n-1} + \cdots + \lambda_p X_{n-p}.$$

This is not necessarily the case. The crucial point concerns the correlation between ξ_n and the past $(X_{n-1}, \ldots, X_{n-p})$. Such prediction problems are the subject of the next section. Here we content ourselves with the preliminary work of determining when a pth order autoregressive process has a moving average representation.

One time unit earlier, (2.9) is

$$X_{n-1} = \lambda_1 X_{n-2} + \lambda_2 X_{n-3} + \cdots + \lambda_p X_{n-p-1} + \xi_{n-1},$$

which inserted back in (2.9) gives

$$\begin{aligned} X_n &= \lambda_1[\lambda_1 X_{n-2} + \cdots + \lambda_p X_{n-p-1} + \xi_{n-1}] \\ &\quad + \lambda_2 X_{n-2} + \cdots + \lambda_p X_{n-p} + \xi_n \\ &= \xi_n + \lambda_1 \xi_{n-1} + (\lambda_1^2 + \lambda_2) X_{n-2} \\ &\quad + \cdots + (\lambda_1 \lambda_{p-1} + \lambda_p) X_{n-p} + \lambda_1 \lambda_p X_{n-p-1}. \end{aligned}$$

Following this procedure m times brings us to

$$\begin{aligned} X_n = \xi_n + \delta_1 \xi_{n-1} + \cdots + \delta_m \xi_{n-m} + \beta_{m1} X_{n-m-1} \\ + \beta_{m2} X_{n-m-2} + \cdots + \beta_{mp} X_{n-m-p}, \end{aligned} \tag{2.10}$$

for certain constants δ_i and β_{mi}. (Each substitution leaves p consecutive X_n's on the right.) Then, substitution of

$$X_{n-m-1} = \xi_{n-m-1} + \lambda_1 X_{n-m-2} + \cdots + \lambda_p X_{n-m-p-1}$$

into (2.10) gives

$$\begin{aligned} X_n = \xi_n + \delta_1 \xi_{n-1} + \cdots + \delta_m \xi_{n-m} + \beta_{m1} \xi_{n-m-1} \\ + (\beta_{m1} \lambda_1 + \beta_{m2}) X_{n-m-2} + \cdots \\ + (\beta_{m1} \lambda_{p-1} + \beta_{mp}) X_{n-m-p} + \beta_{m1} \lambda_p X_{n-m-p-1}, \end{aligned}$$

from which we deduce the recurrence relations

$$\delta_{m+1} = \beta_{m1},$$
$$\beta_{m+1,j} = \beta_{m1}\lambda_j + \beta_{m,j+1}, \qquad j = 1, \ldots, p-1,$$

and

$$\beta_{mp} = \beta_{m1}\lambda_p.$$

Continuation of this procedure leads to the moving average representation

$$X_n = \sum_{k=0}^{\infty} \delta_k \xi_{n-k}, \qquad \delta_0 = 1, \tag{2.11}$$

provided the remainder terms $R_m = (\beta_{m1}X_{n-m-1} + \cdots + \beta_{mp}X_{n-m-p})$ *become negligible in mean square as* $m \to \infty$. We will derive momentarily the proper conditions to insure formula (2.11).

The procedure becomes neater when formulated in a more abstract setting. Let T be the *shift operator* which takes the real sequence $\{x_n\} = (\ldots, x_{-1}, x_0, x_1, \ldots)$ into the sequence $T(\{x_n\}) = (\ldots, x_0, x_1, x_2, \ldots)$ (i.e., x_1 is the zeroth coordinate). Each time index has been advanced by one. Of course, T^{-1} works in the opposite direction:

$$T^{-1}(\{x_n\}) = (\ldots, x_{-2}, x_{-1}, x_0, \ldots),$$

and

$$T^{-k}(\{x_n\}) = (\ldots, x_{-k-1}, x_{-k}, x_{-k+1}, \ldots).$$

In this notation, (2.9) becomes

$$\{X_n\} = \lambda_1 T^{-1}(\{X_n\}) + \cdots + \lambda_p T^{-p}(\{X_n\}) + \{\xi_n\}$$
$$= \sum_{k=1}^{p} \lambda_k T^{-k}(\{X_n\}) + \{\xi_n\},$$

or

$$\left(I - \sum_{k=1}^{p} \lambda_k T^{-k}\right)(\{X_n\}) = \{\xi_n\},$$

where $I = T^0$ is the identity operator. Now we see that what is needed is an inverse to the operator $\left(I - \sum_{k=1}^{p} \lambda_k T^{-k}\right)$ such that we are permitted to write

$$\{X_n\} = \left(I - \sum_{k=1}^{p} \lambda_k T^{-k}\right)^{-1}(\{\xi_n\}).$$

That is, if

$$X_n = \sum_{i=0}^{\infty} \delta_i \xi_{n-i}, \qquad \text{or} \qquad \{X_n\} = \left(\sum_{i=0}^{\infty} \delta_i T^{-i}\right)(\{\xi_n\}),$$

then the δ_i's are the coefficients in

$$\left(1 - \sum_{k=1}^{p} \lambda_k z^k\right)^{-1} = \sum_{k=0}^{\infty} \delta_k z^k.$$

As such, they can be determined by formal long division. For example, the case $\lambda_1 = \lambda$, $\lambda_k = 0$, for $k \geq 2$, which immediately preceded this general discussion, is solved by

$$\frac{1}{1 - \lambda z} = 1 + \lambda z + \lambda^2 z^2 + \lambda^3 z^3 + \cdots,$$

or

$$X_n = \sum_{j=0}^{\infty} \lambda^j \xi_{n-j},$$

which was obtained in (2.8).

A tedious but straightforward computation reveals that

$$\frac{1}{1 - \lambda_1 z - \cdots - \lambda_p z^p} = 1 + \delta_1 z + \cdots + \delta_m z^m + r_m(z),$$

where

$$r_m(z) = \frac{\beta_{m1} z^{m+1} + \cdots + \beta_{mp} z^{m+p}}{1 - \lambda_1 z - \cdots - \lambda_p z^p}.$$

That is, the formal division yields exactly the same recursion as we obtained earlier through direct substitution. It follows that if $r_m(z) \to 0$ as $m \to \infty$, then each $\beta_{mk} \to 0$, $k = 1, \ldots, p$, and the random remainder term $R_m = \beta_{m1} X_{n-m-1} + \cdots + \beta_{mp} X_{n-m-p}$ vanishes in mean square, so that the infinite moving average representation (2.11) is valid. We give the conditions. The equation

$$x^p - \lambda_1 x^{p-1} - \cdots - \lambda_p = 0 \tag{2.12}$$

has p roots, $x_1, \ldots, x_p$. If $|x_i| < 1$, $i = 1, \ldots, p$, then the roots of

$$1 - \lambda_1 z - \cdots - \lambda_p z^p = 0$$

are $z_i = 1/x_i$, provided $\lambda_p \neq 0$, and $|z_i| > 1$. For any z satisfying $|z| < \min_i |z_i|$, the series

$$\frac{1}{1 - \lambda_1 z - \cdots - \lambda_p z^p} = \frac{1}{\prod_{i=1}^{p} \left(1 - \frac{z}{z_i}\right)} = \prod_{i=1}^{p} \sum_{v=0}^{\infty} \left(\frac{z}{z_i}\right)^v$$

$$= \sum_{r=0}^{\infty} \delta_r z^r,$$

converges absolutely. Hence $r_m(z) \to 0$ as $m \to \infty$ for $|z| < \min_i |z_i|$, and in particular for $|z| = 1$. This implies $\beta_{mk} \to 0$ as $m \to \infty$ for $k = 1, 2, \ldots, p$, and, easily, R_m converges to zero in mean square so that the infinite moving average representation (2.11) holds.

For future reference, we highlight an important conclusion emanating from the preceding analysis.

Remark 2.1. Suppose $\{X_n\}$ is a zero-mean covariance stationary process satisfying

$$X_n = \lambda_1 X_{n-1} + \cdots + \lambda_p X_{n-p} + \xi_n, \qquad n = 0, \pm 1, \ldots,$$

where $\{\xi_n\}$ is a sequence of zero-mean uncorrelated random variables, and $\lambda_1, \ldots, \lambda_p$ are fixed. If every root x_i, $i = 1, \ldots, p$, of

$$x^p - \lambda_1 x^{p-1} - \cdots - \lambda_p = 0$$

has magnitude $|x_i| < 1$, then ξ_n and $(X_{n-1}, \ldots, X_{n-p})$ are uncorrelated. This follows since the terms in $X_{n-k} = \sum_{j=0}^{\infty} \delta_j \xi_{n-k-j}$ $(k \geq 1)$ and ξ_n are uncorrelated by assumption.

To glimpse at the alternative, involving a moving average in the opposite direction, look at

$$X_n = \eta_n + \rho \eta_{n+1} + \rho^2 \eta_{n+2} + \cdots, \tag{2.13}$$

where $|\rho| < 1$ and $\{\eta_n\}$ are zero mean and uncorrelated. Then

$$\rho X_n = X_{n-1} - \eta_{n-1}, \qquad \text{or} \qquad X_n = \lambda X_{n-1} + \xi_n,$$

where $\lambda = 1/\rho$ and $\xi_n = -\rho^{-1}\eta_{n-1}$. Since $\{\xi_n\}$ are uncorrelated, $\{X_n\}$ is an autoregressive process, but $|\lambda| > 1$ precludes a moving average representation of the form (2.11). Our construction of the process in (2.13) is as a moving average in the forward direction, and this exhibits typical behavior when the roots of (2.12) all exceed one in magnitude. When the sizes of the roots of (2.12) lie on both sides of one, a doubly infinite moving average $\sum_{k=-\infty}^{+\infty} \delta_k \xi_{n-k}$ is required to represent $\{X_n\}$. Finally, if any root has magnitude exactly one, the autoregressive process satisfying (2.9) cannot be stationary (except in the trivial case $\xi_n \equiv 0$).

The autoregressive and moving average models may be combined to generate more complex stationary processes. One might suppose

$$X_n = \lambda_1 X_{n-1} + \cdots + \lambda_p X_{n-p} + \eta_n,$$

where

$$\eta_n = \alpha_1 \xi_n + \alpha_2 \xi_{n-1} + \cdots + \alpha_m \xi_{n-m+1},$$

is a moving average. More generally, one can also consider

$$Z_n = X_n + \varepsilon_n,$$

where $\{\varepsilon_n\}$ is an uncorrelated "noise" process. The analysis of these more complex models is quite difficult.

3: Mean Square Error Prediction

A problem that arises in many, many contexts is the problem of predicting the value of a given random variable that will be observed in the future. The newspapers often contain predictions for future values of Gross National Product, employment, consumer prices, and other economic quantities. The problem of control, say process control, implicitly involves prediction. If the predicted output of a process is not satisfactory, a control or change is implemented to bring the process more in line with what is desired.

This section considers the prediction problem in the context of stationary random processes. We suppose we are concerned with a quantity whose values at times $n = \ldots, -2, -1, 0, 1, 2, \ldots$ form a stationary process $\{X_n\}$. We want to predict the value X_{n+1}, or more generally X_{n+k}, based on the values observed in the past, $X_n, X_{n-1}, X_{n-2}, \ldots$. This is the *prediction* or *extrapolation* problem for stationary processes.

GENERAL PREDICTION THEORY

Let X denote the outcome of some "experiment." We suppose that X is unknown, to be observed in the future, and that it is desired to predict the value for X that will occur. Let $\hat{X}$ be our prediction for X. Our prediction error will be $X - \hat{X}$, and it is desired to make this as small as possible, in some sense. A problem arises immediately: What is meant by "small"? In general, $X - \hat{X}$ will be a random quantity, so that we must mean small *on the average*, or that the expected value of some function of the error should be small. Whatever function of the error we choose, it seems reasonable to suppose that it has a minimum at an error of zero and increases as the error deviates from zero. In this section, we measure the performance of a predictor by its *mean squared error* $E[(X - \hat{X})^2] = \|X - \hat{X}\|^2$. This criterion reflects a reasonable requirement that large errors are more serious than small ones. However, perhaps even more important is the mathematical tractability of this criterion which, as we will see, enables a general theory to be developed that leads to explicit formulas for optimal predictors in a number of specific situations.

Thus our problem is to find a predictor that minimizes over all

predictors $\hat{X}$ the mean square error $E[(X-\hat{X})^2]$. To complete the description we must specify the class of allowable predictors over which the minimization takes place. In general, this class is determined by the knowledge concerning the experiment or the distribution of X that is available. To bring this explicitly into the general formulation, let us denote the class of allowable predictors by **H**.

Examples. (i) We suppose X, the outcome of an experiment, has known mean μ and variance σ^2. The best predictor for X, in the sense of mean square error and in the absence of further information, is $\hat{X}=\mu$. Since no further information concerning X is available, we take our space of predictors **H** to be the set of real numbers. We may choose any real number a as our predicted value for X. We compute

$$\begin{aligned}E[(X-\hat{X})^2] &= E[(X-\mu+\mu-\hat{X})^2] \\ &= E[(X-\mu)^2]+2E[(X-\mu)(\mu-\hat{X})]+E[(\mu-\hat{X})^2].\end{aligned}$$

Then $E[(X-\mu)^2]=\sigma^2$, and since $\hat{X}$ is a fixed real number, $E[(\mu-\hat{X})^2]=(\mu-\hat{X})^2$ and $E[(X-\mu)(\mu-\hat{X})]=(\mu-\hat{X})E[X-\mu]=0$. Thus,

$$E[(X-\hat{X})^2]=\sigma^2+(\mu-\hat{X})^2.$$

It follows immediately that the mean square error is minimized by setting $\hat{X}=\mu$. Thus, in the absence of further information, the minimum mean square error predictor for a random variable X is its mean $\mu=E[X]$.

(ii) Now let X, Y be jointly distributed random variables having finite variances and a known joint distribution. We suppose that X is to be predicted from an observation on Y. Common examples are the prediction of tensile strength of a specimen of steel from a reading of its hardness and the prediction of the true value of a physical quantity such as pressure, temperature, etc., based on a measurement subject to random error.

Since we suppose Y is observable and no further information on the outcome for X is available (other than knowing the joint distribution), we allow any function $\hat{X}=f(Y)$ having finite variance as a predictor.

We might argue that, after observing $Y=y$, we are in the same situation as Example (i) except that in this second case the appropriate distribution is the conditional distribution for X given $Y=y$. Therefore the minimum mean square error predictor would be the mean of X computed under this conditional distribution, $\mu_{X|Y}=E[X|Y]$. This is indeed the case. We compute

$$\begin{aligned}&E[(X-\hat{X})^2] \\ &= E[(X-\mu_{X|Y})^2]+2E[(X-\mu_{X|Y})(\mu_{X|Y}-\hat{X})]+E[(\mu_{X|Y}-\hat{X})^2].\end{aligned}$$

We show that the expectation of the middle term on the right vanishes by conditioning on Y, and recalling that $\mu_{X|Y}$ and $\hat{X}$ are functions of Y. We have

$$\begin{aligned} E[(X-\mu_{X|Y})(\mu_{X|Y}-\hat{X})] &= E\{E[(X-\mu_{X|Y})(\mu_{X|Y}-\hat{X})|Y]\} \\ &= E\{(\mu_{X|Y}-\hat{X})E[(X-\mu_{X|Y})|Y]\} \\ &= 0, \end{aligned}$$

since $E[(X-\mu_{X|Y})|Y] = E[X|Y] - \mu_{X|Y} = 0$. Thus

$$E[(X-\hat{X})^2] = E[(X-\mu_{X|Y})^2] + E[(\mu_{X|Y}-\hat{X})^2],$$

and the right-hand side is minimized by setting $\hat{X} = E[X|Y]$.

(iii) The computation of $\hat{X} = E[X|Y]$ requires full knowledge of the joint distribution of X, Y. Even when this knowledge is available, the resulting formulas are often too complicated to be of practical value. In the study of covariance stationary processes we assume knowledge of the first two moments of the process only, and no further information on the joint distribution is assumed available. Therefore, it is desirable to have a prediction theory that both leads to simple predictor formulas and requires knowledge of the first two moments only. This is the theory of *linear predictors*. Let X, Y be jointly distributed random variables having known means μ_X, μ_Y, respectively, variances σ_X^2, σ_Y^2, respectively, and covariance $\sigma_{X,Y} = E[(X-\mu_X)(Y-\mu_Y)]$. We allow as predictors only those formulas that are linear functions of Y, say $\hat{X} = a + bY$, or, what is equivalent and more convenient, $\hat{X} = a + b(Y-\mu_Y)$, where a and b are arbitrary real numbers. Since the class of allowable predictors is smaller in this example than in Example (ii), the resulting minimum mean square prediction error cannot be smaller and quite possibly might be larger. However the predictor formula, being linear, is simple, and we will show that the optimal coefficients can be determined without knowing the full joint distribution, but knowing only the given moments. The assertion is that the optimal linear predictor of X based on Y is

$$\hat{X}^* = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(Y-\mu_Y).$$

We proceed to validate this claim. Let

$$\hat{X} = a + b(Y-\mu_Y),$$

and note that

$$\hat{X}^* - \hat{X} = a' + b'(Y-\mu_Y),$$

where $a' = \mu_X - a$ and $b' = (\sigma_{XY}/\sigma_Y^2) - b$. As before, we compute

$$E[(X - \hat{X})^2]$$
$$= E[(X - \hat{X}^*)^2] + 2E[(X - \hat{X}^*)(\hat{X}^* - \hat{X})] + E[(\hat{X}^* - \hat{X})^2].$$

We again determine that the second term on the right vanishes. Indeed,

$$\begin{aligned} E[(X - \hat{X}^*)(\hat{X}^* - \hat{X})] &= E\left[\left\{(X - \mu_X) - \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \mu_Y)\right\}\right. \\ &\qquad \left. \times \{a' + b'(Y - \mu_Y)\}\right] \\ &= a'E\left[(X - \mu_X) - \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \mu_Y)\right] \\ &\quad + b'E\left[(X - \mu_X)(Y - \mu_Y) - \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \mu_Y)^2\right] \\ &= 0 + b'\left(\sigma_{XY} - \frac{\sigma_{XY}}{\sigma_Y^2}\,\sigma_Y^2\right) \\ &= 0. \end{aligned}$$

Thus we have

$$E[(X - \hat{X})^2] = E[(X - \hat{X}^*)^2] + E[(\hat{X}^* - \hat{X})^2],$$

and the right-hand side is minimized over linear predictors $\hat{X}$ by setting $\hat{X} = \hat{X}^*$. Thus $\hat{X}^*$ is the minimum mean square error linear predictor, as stated.

There is an important special case in which the restriction to linear predictor formulas results in no loss of prediction efficiency. If X and Y have a joint normal distribution, then, as shown in Chapter 1, the conditional mean of X given $Y = y$ is

$$E[X|\,Y = y] = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \mu_Y).$$

We observe that this best predictor is linear in y. Thus, if X and Y have a joint normal distribution, the minimum mean square error *linear* predictor of X given Y is, in fact, the minimum mean square error predictor.

A pattern emerges in these examples. In each case, to show that a given predictor $\hat{X}^*$ was optimal, the crux of the matter lay in showing that a cross product, $E[(X - \hat{X}^*)(\hat{X}^* - \hat{X})]$, vanished. We formalize the pattern into *the prediction theorem for minimum mean square error predictors.*

Theorem 3.1. *Let X be a random variable having a finite second moment and let* **H** *be a space of allowable predictors $\hat{X}$, that is, a set of random variables $\hat{X}$ having finite second moments. Assume that* **H** *is a linear space in the sense that a $\hat{X}_1 + \hat{X}_2$ is allowed as a predictor whenever $\hat{X}_1$ and $\hat{X}_2$ are predictors and a is a real number. Then:*

(i) *A predictor $\hat{X}^*$ has minimum mean square error if and only if $E[(X - \hat{X}^*)U] = 0$ for every predictor U;*

(ii) *If a minimum mean square error predictor exists, it is unique in the sense of mean square distance. That is, if $\hat{X}_1^*$ and $\hat{X}_2^*$ are minimum mean square error predictors, then $E[(\hat{X}_1^* - \hat{X}_2^*)^2] = 0$;*

(iii) *To assure the existence of a minimum mean square error predictor, it is sufficient to assume that* **H** *is closed in the sense that, if $\hat{X}_n$; $n = 0, 1, \ldots$, is a sequence of predictors converging in mean square to a random variable $\hat{X}$, then $\hat{X}$ is also allowed as a predictor.*

Proof. (i) For the first part we suppose that $\hat{X}^*$ is a predictor for which

$$E[(X - \hat{X}^*)U] = 0,$$

for every predictor U. We will show that $\hat{X}^*$ represents the minimum mean square prediction error. Let $\hat{X}$ be an arbitrary allowable predictor and compute

$$E[(X - \hat{X})^2] = E[(X - \hat{X}^*)^2] + 2E[(X - \hat{X}^*)(\hat{X}^* - \hat{X})] + E[(\hat{X}^* - \hat{X})^2].$$

Set $U = \hat{X}^* - \hat{X}$. Then U is a predictor since both $\hat{X}^*$ and $\hat{X}$ are and any linear combination of predictors is an allowable predictor. Thus $E[(X - \hat{X}^*)(\hat{X}^* - \hat{X})] = E[(X - \hat{X}^*)U] = 0$, and

$$E[(X - \hat{X})^2] = E[(X - \hat{X}^*)^2] + E[(\hat{X}^* - \hat{X})^2].$$

Hence

$$E[(X - \hat{X})^2] \geq E[(X - \hat{X}^*)^2],$$

and the mean square error for $\hat{X}^*$ is smaller than that for any other predictor $\hat{X}$. Thus $\hat{X}^*$ achieves minimum mean square prediction error.

On the other hand, suppose for a given predictor $\hat{X}^*$ that $E[(X - \hat{X}^*)U] = a \neq 0$ for some predictor U. We will show that

$$\hat{X} = \hat{X}^* + \frac{a}{E[U^2]}\, U \tag{3.1}$$

provides a smaller mean square error, and hence $\hat{X}^*$ cannot be optimal.

This will complete the proof of (i). First note that $\hat{X}$ is a predictor since it is a linear combination of predictors. Rewriting (3.1) as

$$\hat{X} - \hat{X}^* = \frac{a}{E[U^2]}\, U$$

and inserting this twice in

$$\begin{aligned} E[(X-\hat{X})^2] &= E[\{(X-\hat{X}^*) - (\hat{X}-\hat{X}^*)\}^2] \\ &= E[(X-\hat{X}^*)^2] - 2E[(X-\hat{X}^*)(\hat{X}-\hat{X}^*)] \\ &\quad + E[(\hat{X}-\hat{X}^*)^2], \end{aligned}$$

we get

$$\begin{aligned} E[(X-\hat{X})^2] &= E[(X-\hat{X}^*)^2] - 2\frac{a}{E[U^2]}E[(X-\hat{X}^*)U] \\ &\quad + \frac{a^2}{\{E[U^2]\}^2}E[U^2] \\ &= E[(X-\hat{X}^*)^2] - \frac{a^2}{E[U^2]} \\ &< E[(X-\hat{X}^*)^2]. \end{aligned}$$

We conclude that $\hat{X}^*$ cannot provide minimum mean square error since we have exhibited a predictor $\hat{X}$ with smaller error.

(ii) To demonstrate that the minimum mean square error predictor is unique, let us suppose both $\hat{X}_1^*$ and $\hat{X}_2^*$ have minimum mean square error. Then $E[(X-\hat{X}_i^*)U] = E[XU - \hat{X}_i^*U] = E[XU] - E[\hat{X}_i^*U] = 0$, for $i = 1, 2$ and all predictors U. Thus

$$E[\hat{X}_1^*U] = E[XU] = E[\hat{X}_2^*U],$$

and

$$E[(\hat{X}_1^* - \hat{X}_2^*)U] = 0,$$

for all predictors U. We choose $U = \hat{X}_1^* - \hat{X}_2^*$ to conclude

$$E[(\hat{X}_1^* - \hat{X}_2^*)^2] = 0. \tag{3.2}$$

Thus minimum mean square error predictors are unique in the sense that the mean square difference between them is zero. Using Chebyshev's inequality one can show, as was done earlier in this chapter, that (3.2) implies

$$\Pr\{\hat{X}_1^* \neq \hat{X}_2^*\} = 0.$$

(iii) Let d be the infimum of the mean square errors of predictors in $\mathbf{H}$,

$$d = \inf\{E[|X - \hat{X}|^2] : \hat{X} \in \mathbf{H}\}.$$

We may select a sequence $\{\hat{X}_n\}$ of predictors from $\mathbf{H}$ whose mean square errors approach d,

$$\lim_{n\to\infty} E[|X - \hat{X}_n|^2] = d.$$

For an arbitrary m, n, we apply the parallelogram law with

$$Y = X - \hat{X}_n, \qquad Z = X - \hat{X}_m,$$

to conclude

$$E[|\hat{X}_m - \hat{X}_n|^2] + E[|2X - \hat{X}_n - \hat{X}_m|^2] = 2E[|X - \hat{X}_n|^2] + 2E[|X - \hat{X}_m|^2]. \tag{3.3}$$

Since $\frac{1}{2}(\hat{X}_n + \hat{X}_m)$ is a predictor, its mean square error must exceed d. Hence

$$E[|2X - \hat{X}_n - \hat{X}_m|^2] = 4E[|X - \tfrac{1}{2}(\hat{X}_n + \hat{X}_m)|^2] \geq 4d.$$

Inserting this in (3.3) gives

$$E[|\hat{X}_n - \hat{X}_m|^2] + 4d \leq 2E[|\hat{X}_n - X|^2] + 2E[|\hat{X}_m - X|^2].$$

We subtract $4d$ from both sides and let m, n increase indefinitely to conclude

$$\begin{aligned}\lim_{m,n\to\infty} E[|\hat{X}_m - \hat{X}_n|^2] &\leq 2 \lim_{n\to\infty} E[|\hat{X}_n - X|^2] + 2 \lim_{m\to\infty} E[|\hat{X}_m - X|^2] - 4d \\ &= 0.\end{aligned}$$

Since the Cauchy criterion of mean square convergence is satisfied, we know there exists a random variable, call it $\hat{X}^*$, for which $\hat{X}_n \to \hat{X}^*$(ms). By assumption, $\hat{X}^*$ is a predictor. Then $X - \hat{X}_n \to X - \hat{X}^*$(ms), and by the continuity of the mean square in mean square convergence

$$E[|X - \hat{X}^*|^2] = \lim_{n\to\infty} E[|X - \hat{X}_n|^2] = d.$$

Thus $\hat{X}^*$ is a minimum mean square error predictor and the proof is complete. ■

In Examples (i)–(iii) our approach was to verify that given predictors were optimal. Let us return to those examples and show how, with the aid of the previous theorem, we might derive the optimal predictors. We use part (i), the necessary and sufficient condition for a predictor to have

minimum mean square error. In Example (i), **H** is the set of real numbers, and we desire a real number $\hat{X}^*$ for which

$$E[(X-\hat{X}^*)u]=0, \tag{3.4}$$

for all real numbers u. It clearly suffices to take $u=1$. Then

$$E[(X-\hat{X}^*)]=0,$$

and, since $\hat{X}^*$ is nonrandom,

$$\hat{X}^*=E[\hat{X}^*]=E[X]=\mu.$$

Since $\hat{X}^*$ so chosen satisfies (3.4), we know $\hat{X}^*$ is the (unique) minimum mean square error predictor.

For Example (ii), **H** is the set of all functions $\hat{X}=f(Y)$ of Y having finite second moments. The necessary and sufficient condition for $\hat{X}^*$, a function of Y, to be optimal is that

$$E[(X-\hat{X}^*)f(Y)]=0,$$

for all functions of Y having finite second moments, or

$$E[Xf(Y)]=E[\hat{X}^*f(Y)].$$

This is the defining property for $\hat{X}^*$ to be the conditional expectation of X given Y (see Chapter 1). Thus, $\hat{X}^*=E[X|Y]$ is the (unique in the sense of mean square distance) optimal predictor.

The condition in Example (iii) is

$$E[(X-\hat{X}^*)U]=0, \tag{3.5}$$

whenever U is a linear function of Y. We write

$$U=a+b(Y-\mu_Y),$$
$$\hat{X}^*=a^*+b^*(Y-\mu_Y),$$

so that (3.5) becomes

$$E[\{X-(a^*+b^*(Y-\mu_Y))\}\times\{a+b(Y-\mu_Y)\}]=0.$$

This must hold for all choices of a and b. In particular, let us choose first $a=1$ and $b=0$, and then $a=0$ and $b=1$, to conclude

$$E[X-(a^*+b^*(Y-\mu_Y))]=0,$$

or

$$E[X]=a^*+b^*E[Y-\mu_Y]=a^*,$$

and

$$E[\{X-a^*-b^*(Y-\mu_Y)\}(Y-\mu_Y)]=0,$$

or

$$E[(X-\mu_X)(Y-\mu_Y)] = b^*E[(Y-\mu_Y)^2],$$
$$\sigma_{XY} = b^*\sigma_Y^2.$$

Thus $a^* = \mu_X$ and $b^* = \sigma_{XY}/\sigma_Y^2$, so

$$\hat{X}^* = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(Y-\mu_Y).$$

Note. Examining some parallels in finite-dimensional Euclidean geometry leads to a better understanding of why it is possible to develop a nice theory of prediction under the mean square error criterion. Recall that a (real) *vector space* is a set of elements x, y with associated operations of vector addition and scalar multiplication such that $ax + by$ is a vector whenever x and y are vectors, and a and b are real numbers. An *inner product* in a vector space is a real-valued function denoted (x, y) of pairs of vectors x, y satisfying $(x, y) = (y, x)$, $(a_1x_1 + a_2x_2, y) = a_1(x_1, y) + a_2(x_2, y)$, and $(x, x) > 0$ if x is not the zero vector. In an inner product space we can define the *norm* or length of a vector x by $\|x\| = (x, x)^{1/2}$. A *Hilbert space* is an inner product space that is complete in the metric determined by the norm. This means that if $\{x_n\}$ is a sequence of vectors for which $\lim_{m,n\to\infty}\|x_n - x_m\| = 0$, then there exists a vector x for which $\lim_{n\to\infty}\|x_n - x\| = 0$.

Under the full assumptions of Theorem 1, **H**, the space of predictors $\hat{X}$, $\hat{Y}$, is a Hilbert space where the inner product is $(\hat{X}, \hat{Y}) = E[\hat{X}\hat{Y}]$. The space **H′** consisting of all random variables of the form $aX + b\hat{X}$, where a and b are real and $\hat{X}$ is in **H**, i.e., a predictor, is also a Hilbert space. The mean square error is the square of the norm or length of the difference vector $X - \hat{X}$. That is, it is the square of the distance between X and $\hat{X}$. In Hilbert space terminology, our problem is to find a vector $\hat{X}^*$ in **H** that is nearest to the vector X in **H′**.

Hilbert space is a generalization of d-dimensional Euclidean space having vectors $x = (x_1, \ldots x_d)$, $y = (y_1, \ldots, y_d)$. The inner product is $(x, y) = \sum x_i y_i$, and the norm is the Euclidean distance $\|x\| = \{\sum x_i^2\}^{1/2}$. The cosine of the angle between two vectors x, y is often defined as $(x, y)/(\|x\|\,\|y\|)$. In particular, two vectors x, y are perpendicular if and only if $(x, y) = 0$.

The problem is to find a vector $\hat{x}^*$ in a subspace **H** of a space **H′** that is nearest to a vector x in **H′**. Theorem 1 states that $\hat{x}^*$ is nearest to x if and only if the difference $x - \hat{x}^*$ is perpendicular to all vectors u in **H**. Figure 1 will help guide the geometric intuition.

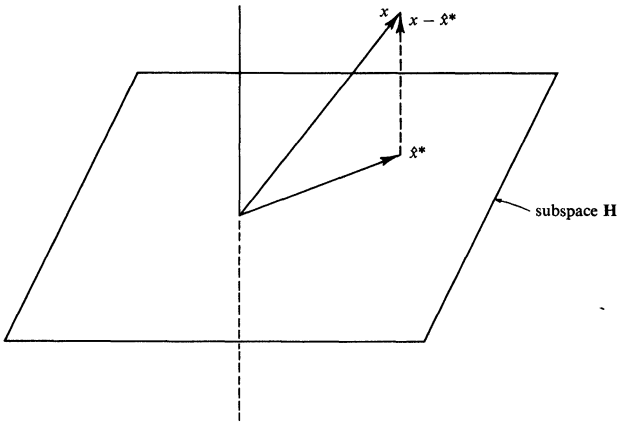


FIG. 1. Diagram showing that a vector $\hat{x}^*$ in a subspace **H** is nearest to a vector x if and only if the difference $x - \hat{x}^*$ is perpendicular to all vectors in **H**.

4: Prediction of Covariance Stationary Processes

Let X_n, $n = 0, \pm 1, \pm 2, \ldots$, be a covariance stationary process having a known covariance function $R(v)$, $v = 0, \pm 1, \ldots$. (Where X_n is a real valued process then $R(-v) = R(v)$.) The mean of the process is constant, assumed known, and thus, after changing the origin if necessary, we may assume the mean is zero. Let us consider the problem of finding the minimum mean square error linear predictor of X_n based on the *finite* past $X_{n-1}, X_{n-2}, \ldots, X_{n-p}$. Thus consider predictors of the form

$$\hat{X}_n = \alpha_1 X_{n-1} + \alpha_2 X_{n-2} + \cdots + \alpha_p X_{n-p}.$$

Using our criterion for optimality, we know that a predictor

$$\hat{X}_n^* = \alpha_1^* X_{n-1} + \alpha_2^* X_{n-2} + \cdots + \alpha_p^* X_{n-p}$$

has minimum mean square error if and only if, for all predictors

$$U = u_1 X_{n-1} + u_2 X_{n-2} + \cdots + u_p X_{n-p},$$

we have

$$E[(X_n - \hat{X}_n^*)U] = 0.$$

Let us consider the p particular choices for U given by

$$U_i = X_{n-i}, \qquad \text{for} \quad i = 1, \ldots, p.$$

Since any predictor U may be written as $U = u_1 U_1 + \cdots + u_p U_p$,

$$E[(X_n - \hat{X}_n^*)U] = 0, \qquad \text{for all} \quad U,$$

if and only if

$$E[(X_n - \hat{X}_n^*)U_i] = 0, \qquad \text{for} \quad i = 1, \ldots, p. \tag{4.1}$$

Substituting in (4.1) we get

$$E[(X_n - \{\alpha_1^* X_{n-1} + \alpha_2^* X_{n-2} + \cdots + \alpha_p^* X_{n-p}\})X_{n-i}] = 0,$$

or

$$E[X_n X_{n-i}] = \alpha_1^* E[X_{n-1}X_{n-i}] + \cdots + \alpha_p^* E[X_{n-p}X_{n-i}].$$

Since these cross products are the covariances, we have

$$R(i) = \alpha_1^* R(i-1) + \cdots + \alpha_p^* R(i-p), \qquad i = 1, \ldots, p.$$

Thus $\alpha^* = (\alpha_1^*, \ldots, \alpha_p^*)$ may be taken as any vector satisfying the p linear equations

$$\begin{aligned}
R(1) &= \alpha_1^* R(0) + \alpha_2^* R(1) + \alpha_3^* R(2) + \cdots + \alpha_p^* R(p-1),\\
R(2) &= \alpha_1^* R(1) + \alpha_2^* R(0) + \alpha_3^* R(1) + \cdots + \alpha_p^* R(p-2),\\
R(3) &= \alpha_1^* R(2) + \alpha_2^* R(1) + \alpha_3^* R(0) + \cdots + \alpha_p^* R(p-3),\\
&\vdots\\
R(p) &= \alpha_1^* R(p-1) + \alpha_2^* R(p-2) + \alpha_3^* R(p-3) + \cdots + \alpha_p^* R(0).
\end{aligned} \tag{4.2}$$

Part (iii) of the basic prediction theorem asserts there is a minimum mean square error predictor in this case. (One can show that the required hypothesis is satisfied.) Thus, there is at least one solution α^* to (4.2), and there may be many. However, by the uniqueness part of the basic prediction theorem, all solutions lead to the same $\hat{X}_n^*$. As an example, consider the extreme $\{Z_n\}$ process of Section 1 in which $Z_n = Z$, $n = 0, \pm 1, \ldots$, and $R(v) = \sigma^2$ for all $v = 0, \pm 1, \ldots$. Then any α^* vector satisfying $\alpha_1^* + \cdots + \alpha_p^* = 1$ will solve (4.2). All such α^* vectors lead to $\hat{Z}_n^* = Z$, however.

We can compute the minimum mean square error σ_p^2 by

$$\begin{aligned}
\sigma_p^2 &= E[|X_n - \hat{X}_n^*|^2]\\
&= E[(X_n - \hat{X}_n^*)(X_n - \hat{X}_n^*)]\\
&= E[(X_n - \hat{X}_n^*)X_n] - E[(X_n - \hat{X}_n^*)\hat{X}_n^*].
\end{aligned}$$

The second term vanishes by the necessary and sufficient condition for a predictor to be optimal. Keeping the first term,

$$\begin{aligned}
\sigma_p^2 &= E[X_n^2] - E[\hat{X}_n^* X_n]\\
&= R(0) - E\left[\sum_{k=1}^{p} \alpha_k^* X_{n-k} X_n\right]\\
&= R(0) - \sum_{k=1}^{p} \alpha_k^* R(k).
\end{aligned}$$

Let us now consider the problem of linearly predicting X_n based on the entire past $X_{n-1}, X_{n-2}, \ldots$. We allow as predictors all random variables of the form

$$\hat{X}_n = \alpha_1 X_{n-1} + \cdots + \alpha_p X_{n-p},$$

where p is any positive integer and $\alpha_1, \ldots, \alpha_p$ are real numbers, and also all limits, in the mean square sense, of such random variables. The space **H** of such predictors satisfies all conditions in the basic prediction theorem. Clearly every random variable of the form

$$\hat{X}_n = \sum_{k=1}^{\infty} \alpha_k X_{n-k} \tag{4.3}$$

is a predictor, provided the α's are chosen so that the infinite sum converges in the mean square sense. By applying the Cauchy criterion to the sequence of partial sums, it can be be shown that the infinite sum will converge whenever $\sum_{k,l} \alpha_k \alpha_l R(|k-l|) < \infty$. No matter how counter-intuitive it may seem, it is unfortunately true that *not* every limit of finite predictors can be represented in the manner of (4.3).

Examples are easy. Let $Z, \ldots, \zeta_{-1}, \zeta_0, \zeta_1, \ldots$ be independent identically distributed random variables having zero mean and unit variance. Let $X_n = Z + \zeta_n$. By the mean square law of large numbers,

$$Z = Z + \lim_{m\to\infty} \frac{1}{m}(\zeta_n + \cdots + \zeta_{n-m+1}) = \lim_{m\to\infty} \frac{1}{m}(X_n + \cdots + X_{n-m+1}),$$

so that Z is an allowable predictor, given the entire past, and clearly is the best predictor. But Z cannot be represented in the form of (4.3). However, we know a minimal mean square error predictor $\hat{X}^*$ exists. In most situations of practical interest, such a predictor will have a representation in the form

$$\hat{X}_n^* = \sum_{k=1}^{\infty} \alpha_k^* X_{n-k}, \tag{4.4}$$

and thus it is worthwhile to examine this case in some detail. Since $E[(X_n - \hat{X}_n^*)U] = 0$ for every U that is a finite linear combination of variables in $X_{n-1}, X_{n-2}, \ldots$, we must have

$$E[(X_n - \hat{X}_n^*)X_{n-k}] = 0, \qquad k = 1, 2, \ldots. \tag{4.5}$$

This then will imply that

$$E[(X_n - \hat{X}_n^*)U] = 0, \tag{4.6}$$

for every U of the form $U = u_1 X_{n-1} + u_2 X_{n-2} + \cdots + u_p X_{n-p}$, and, by continuity of the cross product, (4.6) will hold for all allowable predictors

U. Thus (4.5) provides a necessary and sufficient condition for a predictor of the form (4.4) to be optimal.

Let us apply this criterion to the autoregressive processes of Section 2. Suppose that $\{X_n\}$ is a covariance stationary process satisfying

$$X_n = \lambda_1 X_{n-1} + \cdots + \lambda_p X_{n-p} + \xi_n, \tag{4.7}$$

where $\lambda_1, \ldots, \lambda_p$ are fixed and $\{\xi_n\}$ is a sequence of zero mean uncorrelated random variables having a common variance. Write (4.7) in the form

$$X_n = \hat{X}_n + \xi_n,$$

where

$$\hat{X}_n = \lambda_1 X_{n-1} + \cdots + \lambda_p X_{n-p}.$$

We pointed out in Remark 2.1 that ξ_n is uncorrelated with X_{n-k} for $k \geq 1$, *provided all roots* $x_1, \ldots, x_p$ *of*

$$x^p - \lambda_1 x^{p-1} - \cdots - \lambda_p = 0$$

have magnitude $|x_i| < 1$. It follows that, under this condition, $\hat{X}_n$ is a minimum mean square linear predictor for X_n given the entire past, since

$$E[(X_n - \hat{X}_n)X_{n-k}] = E[\xi_n X_{n-k}] = 0, \qquad k = 1, 2, \ldots,$$

and the condition (4.5) is satisfied.

Return to the general case, and suppose there exists an optimal predictor of the form

$$\hat{X}_n^* = \sum_{k=1}^{\infty} \alpha_k^* X_{n-k}. \tag{4.8}$$

As mentioned earlier, this need not be the case. However since the cross product is continuous in the mean square limit, we substitute into (4.5) and interchange the expectation and sum. The result is

$$R(k) = \sum_{l=1}^{\infty} \alpha_l^* R(k-l), \qquad k = 1, 2, \ldots. \tag{4.9}$$

If these equations have a solution $(\alpha_1^*, \alpha_2^*, \ldots)$ for which the series in (4.8) converges in mean square, then, in fact, (4.8) will define a minimum mean square error predictor.

Example. Suppose $\{X_n, n = 0, \pm 1, \ldots\}$ is a covariance stationary process having the covariance function

$$R(v) = \begin{cases} 1, & v = 0, \\ \lambda/(1+\lambda^2), & |v| = 1, \\ 0, & |v| > 1, \end{cases}$$

where $0 < \lambda < 1$. Equations (4.9) become

$$\lambda = (1 + \lambda^2)\alpha_1^* + \lambda\alpha_2^*,$$
$$0 = \lambda\alpha_{k-1}^* + (1 + \lambda^2)\alpha_k^* + \lambda\alpha_{k+1}^*, \qquad k = 2, 3, \ldots.$$

It is easy to verify that a solution is given by

$$\alpha_k^* = -(-\lambda)^k, \qquad k = 1, 2, \ldots,$$

and the minimum mean square error linear predictor is

$$\hat{X}_n^* = +\lambda X_{n-1} - \lambda^2 X_{n-2} + \lambda^3 X_{n-3} - \cdots.$$

An interesting point is that, although X_n and X_{n-k} are uncorrelated for $k \geq 2$, the optimal linear predictor for X_n involves the entire past of the process.

5: Ergodic Theory and Stationary Processes

An ergodic theorem gives conditions under which an average over time

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

of a stochastic process will converge as the number n of observed periods becomes large. A most important ergodic theorem is the strong law of large numbers which, for independent and identically distributed random variables $X_1, X_2, \ldots$ having finite means $m = E[X_k]$, asserts that the sample averages $\bar{X}_n$ will converge to the mean m on a set of outcomes having probability one. In symbols,

$$\Pr\left\{\lim_{n\to\infty} \bar{X}_n = m\right\} = 1.$$

Stationary processes provide a natural setting for generalizations of the law of large numbers, since for such processes, the mean value is a constant $m = E[X_n]$, independent of time.

Let us consider, for the moment, the problem of estimating an unknown mean value function $m(n) = E[X_n]$ of an arbitrary process. In general, we must take a large number N of separate realizations of the process, say

$$\{X_n^1;\ n = 1, 2, \ldots\},$$
$$\{X_n^2;\ n = 1, 2, \ldots\},$$
$$\vdots$$
$$\{X_n^N;\ n = 1, 2, \ldots\},$$

and then calculate the arithmetic means

$$\bar{X}(n) = \frac{1}{N}\{X_n^1 + \cdots + X_n^N\},$$

which we would use as estimates for $m(n)$. Of course, if $m(n)$ were constant, as would be the case for a stationary process, we might additionally average over n time points and calculate a grand mean

$$\bar{\bar{X}} = \frac{1}{n}\{\bar{X}(1) + \cdots + \bar{X}(n)\}.$$

However, the point remains that, in general, to estimate a mean value of a process, separate independent realizations of the entire process are needed. For comparison, let us consider the same estimation problem for a stationary process $\{X_n\}$ that obeys an ergodic theorem

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \to m, \qquad \text{as} \quad n \to \infty. \tag{5.1}$$

In view of (5.1), we now need observe only a single realization of the process, but over a sufficiently long time duration. We then use $\bar{X}_n$ as our estimate of the constant mean m. The practical value of the ergodic theory for stationary processes is due, to a considerable extent, to this fact. For such processes the mean value (and correlation function) often can be estimated from just a single realization of the process.

In this section we will present two ergodic theorems for stationary processes. Just as there are strong and weak laws of large numbers, there are a variety of ergodic theorems, differing in their assumptions and in the modes of convergence. Our first theorem generalizes the weak law, as it is often stated, to cover covariance stationary processes. Here the natural mode of convergence is in mean square. The second theorem generalizes the strong law and requires strict stationarity. Here the convergence is with probability one.

THE MEAN SQUARE ERGODIC THEOREM

Let $\{X_n\}$ be a covariance stationary process with mean m and covariance function

$$R(v) = E[\{X_{n+v} - m\} \cdot \{X_n - m\}].$$

Let

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n).$$

One form of the weak law of large numbers states that $\bar{X}_n$ converges to m in the mean square sense whenever the sequence $\{X_n\}$ is uncorrelated.

The mean square ergodic theorem draws the same conclusion assuming only that the sequence X_n is asymptotically uncorrelated, in the sense that the covariance $R(v)$ has a Cesaro limit of zero as the lag v increases.

Theorem 5.1. *Suppose $\{X_n\}$ is a covariance stationary process having covariance function $R(v)$. Then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{v=0}^{N-1} R(v) = 0, \tag{5.2}$$

if and only if

$$\lim_{N\to\infty} E[(\bar{X}_N - m)^2] = 0. \tag{5.3}$$

Proof. Since $m = E[X_n]$, we recognize (5.3) as the limit of the variance of $\bar{X}_N$, and (5.2) as the limit of the covariance between $\bar{X}_N$ and X_1. Thus the theorem states that the variance of $\bar{X}_N$ converges to zero if and only if the covariance between X_1 and $\bar{X}_N$ converges to zero.

Let

$$Y_n = X_n - m, \qquad \text{and} \qquad \bar{Y}_N = \frac{1}{N}(Y_1 + \cdots + Y_N).$$

Then, using Schwarz' inequality,

$$\begin{aligned} \left\{\frac{1}{N}\sum_{v=0}^{N-1} R(v)\right\}^2 &= \{E[Y_1\ \bar{Y}_N]\}^2 \\ &\le E[Y_1^2] \cdot E[\bar{Y}_N^2] \\ &= R(0) \cdot E[(\bar{X}_N - m)^2]. \end{aligned}$$

Thus (5.3) entails (5.2). To establish the opposite implication, let us suppose that (5.2) holds. We calculate the variance of

$$\bar{Y}_N = \frac{1}{N}(Y_1 + \cdots + Y_N),$$

obtaining

$$\begin{aligned} E[\bar{Y}_N^2] &= \frac{1}{N^2} E\left[\left(\sum_{k=1}^{N} Y_k\right)^2\right] \\ &= \frac{1}{N^2}\left\{E\left[\sum_{k=1}^{N} Y_k^2 + 2\sum_{k<l} Y_k Y_l\right]\right\} \\ &= \frac{1}{N^2}\left\{NR(0) + 2\sum_{l=1}^{N}\sum_{k=1}^{l-1} R(l-k)\right\} \\ &= \frac{1}{N^2}\left\{2\sum_{l=1}^{N}\sum_{v=0}^{l-1} R(v) - NR(0)\right\}. \end{aligned}$$

Observe that

$$\frac{1}{N} R(0) \to 0, \qquad \text{as} \quad N \to \infty,$$

so that we need concentrate only on the first term. For it we have, for any $M < N$,

$$\frac{2}{N^2} \sum_{l=1}^{N} \sum_{v=0}^{l-1} R(v) = \frac{2}{N^2} \left\{ \sum_{l=1}^{M} \sum_{v=0}^{l-1} R(v) + \sum_{l=M+1}^{N} \sum_{v=0}^{l-1} R(v) \right\}.$$

Let $\varepsilon > 0$ be given and, using assumption (5.2), choose M so that

$$\left| \frac{1}{l} \sum_{v=0}^{l-1} R(v) \right| \leq \varepsilon, \qquad \text{if} \quad l \geq M.$$

Then

$$\left| \frac{2}{N^2} \sum_{l=M+1}^{N} l \times \frac{1}{l} \sum_{v=0}^{l-1} R(v) \right| \leq \frac{2}{N^2} \sum_{l=M+1}^{N} l\varepsilon \leq 2\varepsilon.$$

Thus

$$\left| \frac{2}{N^2} \sum_{l=1}^{N} \sum_{v=0}^{l-1} R(v) \right| \leq \frac{2}{N^2} \left| \sum_{l=1}^{M} \sum_{v=0}^{l-1} R(v) \right| + 2\varepsilon.$$

Let $N \to \infty$. Since M is fixed, the first term on the right vanishes. Since ε is an arbitrary positive number, we must have

$$\lim_{N \to \infty} \frac{2}{N^2} \sum_{l=1}^{N} \sum_{v=0}^{l-1} R(v) = 0,$$

and so $E[\bar{Y}_N^2] \to 0$. This completes the proof. ∎

Using Chebyshev's inequality

$$\Pr\{|\bar{X}_N - m| > \varepsilon\} \leq \frac{E[(\bar{X}_N - m)^2]}{\varepsilon^2}, \qquad \varepsilon > 0,$$

we see that (5.3) implies

$$\lim_{N \to \infty} \Pr\{|\bar{X}_N - m| > \varepsilon\} = 0, \tag{5.4}$$

for all $\varepsilon > 0$. Thus, for sufficiently large N, (5.4) furnishes grounds for believing that the time average $\bar{X}_N$ is approximately m, and thus provides an estimate for this unknown mean.

If the correlation becomes negligible for sufficiently large lags, viz.,

$$\lim_{v\to\infty} R(v) = 0, \tag{5.5}$$

then (5.2) will hold, since for any $M < N$ we may write

$$\frac{1}{N}\sum_{v=0}^{N-1} R(v) = \frac{1}{N}\sum_{v=0}^{M-1} R(v) + \frac{1}{N}\sum_{v=M}^{N-1} R(v),$$

and, in view of (5.5), the second term on the right may be made smaller in absolute value than any preassigned ε by choosing M sufficiently large, and for any fixed M, the first term on the right vanishes as N approaches infinity. Thus (5.5) is a sufficient condition to insure that the time averages $\bar{X}_N$ converge to the mean m.

Suppose $\{X_n\}$ is a covariance stationary process having a mean of zero. The same theorem may be used to obtain conditions under which the sample average covariance

$$\hat{R}_N(v) = \frac{1}{N}\sum_{k=0}^{N-1} X_k X_{k+v} \tag{5.6}$$

will converge to the covariance

$$R(v) = E[X_n X_{n+v}]. \tag{5.7}$$

If we write $W_k = X_k X_{k+v}$, then (5.6) becomes

$$\hat{R}_N(v) = \frac{1}{N}\sum_{k=0}^{N-1} W_k = \bar{W}_N.$$

Thus, if $\{W_k\}$ forms a covariance stationary process for which

$$\lim_{T\to\infty} \frac{1}{T}\sum_{l=0}^{T-1} E[\{W_n - R(v)\}\cdot\{W_{n+l} - R(v)\}] = 0, \tag{5.8}$$

then

$$\lim_{N\to\infty} E[(\hat{R}_N(v) - R(v))^2] = 0.$$

Of course, the conditions required on the covariance of $\{W_k\}$ are conditions on the fourth product moments of the original $\{X_n\}$ process, which limits the general applicability of this result. However, a Gaussian process is determined by its mean and covariance functions, and thus, for a Gaussian

process, (5.8) can be stated in terms of the covariance. Condition (5.8) requires

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=0}^{T-1}E[\{X_nX_{n+v}-R(v)\}\cdot\{X_{n+l}X_{n+l+v}-R(v)\}]=0,$$

which reduces to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=0}^{T-1}\{E[X_nX_{n+v}X_{n+l}X_{n+l+v}]-R(v)^2\}=0.$$

The fourth product moment of joint normally distributed random variables having zero mean is given by

$$E[X_iX_jX_kX_l]=\sigma_{ij}\sigma_{kl}+\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk},$$

where σ_{ij} is the covariance between X_i and X_j. Relevant to this case, we have

$$E[X_nX_{n+v}X_{n+l}X_{n+l+v}]=R(v)^2+R(l)^2+R(l-v)R(l+v).$$

Thus, in the case of a Gaussian process, (5.8) requires

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=0}^{T-1}\{R(l)^2+R(l-v)R(l+v)\}=0. \tag{5.9}$$

For any real numbers a, b, we have $|ab|\leq a^2+b^2$. Thus

$$|R(l-v)R(l+v)|\leq R(l-v)^2+R(l+v)^2.$$

Now

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=0}^{T-1}R(l)^2=0 \tag{5.10}$$

implies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=0}^{T-1}R(l\pm v)^2=0,$$

since the two sums differ by at most a finite number of terms. In view of the inequality, (5.9) will hold provided only that (5.10) is assumed. We have thus demonstrated the following theorem.

Theorem 5.2. *Let $\{X_n\}$ be a Gaussian covariance stationary process having covariance function $R(v)$ and a mean of zero. If*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{v=0}^{T-1}R(v)^2=0,$$

then, for any fixed $v = 0, \pm 1, \ldots,$

$$\lim_{T\to\infty} E[|\hat{R}_T(v) - R(v)|^2] = 0,$$

where $\hat{R}_T(v)$ *is the sample covariance function*

$$\hat{R}_T(v) = \frac{1}{T}\sum_{l=0}^{T-1} X_l X_{l+v}.$$

As a last topic before proceeding to the strong convergence theorem we will prove a general mean square ergodic theorem.

Theorem 5.3. *Let* $\{X_n\}$ *be a covariance stationary process and let*

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

define the sequence of time averages. Then there exists a random variable $\bar{X}$ *that is the mean square limit of* $\{\bar{X}_n\}$,

$$\lim_{n\to\infty} \|\bar{X}_n - \bar{X}\| = 0.$$

Proof. Before commencing the proof let us remark that, in general, the limit $\bar{X}$ will be random. Theorem 5.1 gives the necessary and sufficient additional assumptions needed if it is to be concluded that $\bar{X}$ is, in fact, not random, but is the constant $m = E[X_1]$.

According to the Cauchy criterion for mean square convergence, to prove the theorem it suffices to show

$$\lim_{m,n\to\infty} \|\bar{X}_n - \bar{X}_m\| = 0.$$

Set

$$\mu_N = \inf\|\lambda_1 X_1 + \cdots + \lambda_N X_N\|,$$

where the infimum is over all nonnegative λ_i satisfying $\lambda_1 + \cdots + \lambda_N = 1$. Notice that $\mu_{N+1} \leq \mu_N$, and set

$$\mu = \lim_{N\to\infty} \mu_N = \inf \mu_N.$$

Suppose, for the moment, that $m < n$ and calculate as follows:

$$\|\bar{X}_m + \bar{X}_n\| = \left\|\left(\frac{1}{m}+\frac{1}{n}\right)X_1 + \cdots + \left(\frac{1}{m}+\frac{1}{n}\right)X_m + \frac{1}{n}X_{m+1} + \cdots + \frac{1}{n}X_n\right\|$$
$$= 2\|\lambda_1 X_1 + \cdots + \lambda_n X_n\| \geq 2\mu,$$

where

$$\lambda_i = \frac{1}{2}\left(\frac{1}{m} + \frac{1}{n}\right), \qquad i = 1, \ldots, m,$$

$$\lambda_i = \frac{1}{2n}, \qquad i = m+1, \ldots, n.$$

Of course, the same inequality obtains if $m > n$. Now, the key to the proof of the theorem is to show

$$\lim_{n\to\infty} \|\bar{X}_n\| = \mu, \tag{5.11}$$

because, if we can do this, from the parallelogram law and the inequality just shown,

$$\begin{aligned}\|\bar{X}_n - \bar{X}_m\|^2 &= 2\|\bar{X}_n\|^2 + 2\|\bar{X}_m\|^2 - \|\bar{X}_n + \bar{X}_m\|^2 \\ &\leq 2\|\bar{X}_n\|^2 + 2\|\bar{X}_m\|^2 - (2\mu)^2 \\ &\leq 2\{|\,\|\bar{X}_n\|^2 - \mu^2| + |\,\|\bar{X}_m\|^2 - \mu^2|\},\end{aligned}$$

and (5.11) implies that the right-hand side vanishes.

Thus we concentrate on verifying (5.11). Let a positive ε be given and choose N and $\lambda_1, \ldots, \lambda_N$ satisfying

$$\|\lambda_1 X_1 + \cdots + \lambda_N X_N\| \leq \mu + \varepsilon,$$

where, of course, $\lambda_i \geq 0$ and $\lambda_1 + \cdots + \lambda_N = 1$. Let

$$Y_k = \lambda_1 X_k + \cdots + \lambda_N X_{k+N-1}.$$

Then $\{Y_k\}$ is a covariance stationary process and

$$\|Y_k\| \leq \mu + \varepsilon, \qquad k = 1, 2, \ldots.$$

We compute

$$\begin{aligned}\bar{Y}_n &= \frac{1}{n}(Y_1 + \cdots + Y_n) \\ &= \frac{1}{n}(\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_N X_N \\ &\quad + \lambda_1 X_2 + \lambda_2 X_3 + \cdots + \lambda_N X_{N+1} + \cdots \\ &\quad + \lambda_1 X_n + \lambda_2 X_{n+1} + \cdots + \lambda_N X_{N+n-1}) \\ &= \lambda_1 \bar{X}_{1,n} + \lambda_2 \bar{X}_{2,n} + \cdots + \lambda_N \bar{X}_{N,n}\end{aligned}$$

where

$$\bar{X}_{k,n} = \frac{1}{n}(X_k + X_{k+1} + \cdots + X_{k+n-1}).$$

Since $\bar{X}_n = \bar{X}_{1,n}$ and $\lambda_1 - 1 = -(\lambda_2 + \cdots + \lambda_N)$, we may write

$$\bar{Y}_n - \bar{X}_n = \lambda_2(\bar{X}_{2,n} - \bar{X}_{1,n}) + \cdots + \lambda_N(\bar{X}_{N,n} - \bar{X}_{1,n}).$$

Then, using the triangle inequality

$$\|\bar{Y}_n - \bar{X}_n\| \leq \lambda_2 \|\bar{X}_{2,n} - \bar{X}_{1,n}\| + \cdots + \lambda_N \|\bar{X}_{N,n} - \bar{X}_{1,n}\|.$$

But

$$\begin{aligned}
\|\bar{X}_{k,n} - \bar{X}_{1,n}\| &= \frac{1}{n} \|(X_k + \cdots + X_{k+n-1}) - (X_1 + \cdots + X_n)\| \\
&= \frac{1}{n} \|X_{n+1} + \cdots + X_{n+k-1} - X_1 - \cdots - X_{k-1}\| \\
&\leq \frac{1}{n} \{\|X_{n+1}\| + \cdots + \|X_{n+k-1}\| + \|X_1\| + \cdots + \|X_{k-1}\|\} \\
&= \frac{2k-2}{n} \|X_1\|, \qquad \text{for} \quad k = 2, \ldots, N.
\end{aligned}$$

Thus

$$\begin{aligned}
\|\bar{Y}_n - \bar{X}_n\| &\leq \sum_{k=2}^{N} \lambda_k \frac{2k}{n} \|X_1\| \\
&\leq \frac{2N\|X_1\|}{n}.
\end{aligned}$$

It follows that

$$\lim_{n \to \infty} \|\bar{Y}_n - \bar{X}_n\| = 0.$$

To conclude the proof of (5.11),

$$\begin{aligned}
\mu \leq \|\bar{X}_n\| &\leq \|\bar{X}_n - \bar{Y}_n\| + \|\bar{Y}_n\| \\
&\leq \|\bar{X}_n - \bar{Y}_n\| + \mu + \varepsilon
\end{aligned}$$

Since $\|\bar{X}_n - \bar{Y}_n\| \to 0$ as $n \to \infty$ and ε is arbitrary, we must have

$$\lim_{n \to \infty} \|\bar{X}_n\| = \mu.$$

This validates (5.11) and completes the proof of the theorem. ■

THE STRONG ERGODIC THEOREM

Theorem 5.4. *Let* $\{X_n; n = 0, 1, 2, \ldots\}$ *be a strictly stationary process having finite mean* $m = E[X_n]$ *Let*

$$\bar{X}_n = \frac{1}{n}(X_0 + \cdots + X_{n-1})$$

be the sample time average. Then, with probability one, the sequence $\{\bar{X}_n\}$ *converges to some limit random variable, denoted* $\bar{X}$. *That is,*

$$\Pr\left\{\lim_{n\to\infty} \bar{X}_n = \bar{X}\right\} = 1.$$

Proof. Let

$$\bar{X}^* = \limsup_{n\to\infty} \bar{X}_n, \qquad \text{and} \qquad \bar{X}_* = \liminf_{n\to\infty} \bar{X}_n.$$

The event that the sequence $\{\bar{X}_n\}$ converges is, of course, the event that $\bar{X}^* = \bar{X}_*$, and the complementary event, the event that $\{\bar{X}_n\}$ does not converge, is the event $\bar{X}^* > \bar{X}_*$. Let K denote this latter event. We want to show

$$\Pr\{K\} = 0.$$

Consider for the moment a particular realization $X_0 = x_0$, $X_1 = x_1$, ..., for which the event K occurs. Then $\bar{X}^* = \bar{x}^* > \bar{X}_* = \bar{x}_*$, and for some rationals $\alpha < \beta$ we must have

$$\bar{x}^* > \beta > \alpha > \bar{x}_*.$$

Since there are only denumerably many pairs of rationals we may let $(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots$ be an enumeration of all such pairs for which $\alpha_k < \beta_k$. Let K_k be the event

$$\bar{X}^* > \beta_k > \alpha_k > \bar{X}_*.$$

If the event K occurs, then for some k, the event K_k must occur. We showed this just above. Conversely, if some K_k occurs, then so does the event K. Thus $K = \bigcup_{k=1}^{\infty} K_k$, and

$$\Pr\{K\} \leq \sum_{k=1}^{\infty} \Pr\{K_k\}.$$

It follows that, to prove $\Pr\{K\} = 0$, it suffices to show that for every k, $\Pr\{K_k\} = 0$. That is, if α and β are arbitrary rational numbers with $\alpha < \beta$, we need only show

$$\Pr\{\bar{X}^* > \beta > \alpha > \bar{X}_*\} = 0.$$

Let A denote the event $\bar{X}^* > \beta > \alpha > \bar{X}_*$ and let $I(A)$ be the indicator random variable of the event A:

$$I(A) = \begin{cases} 1, & \text{if } A \text{ occurs,} \\ 0, & \text{if } A \text{ does not occur.} \end{cases}$$

To show $\Pr\{A\} = E[I(A)] = 0$, we first prove the inequalities

$$E[(X_0 - \beta)I(A)] \geq 0, \tag{5.12}$$

and

$$E[(\alpha - X_0)I(A)] \geq 0. \tag{5.13}$$

These, added together, give

$$E[(\alpha - \beta)I(A)] \geq 0,$$

or

$$(\alpha - \beta)\Pr\{A\} \geq 0.$$

Since $\alpha < \beta$, this will imply $\Pr\{A\} = 0$, which will complete the proof. Our objective will be to establish the inequalities (5.12) and (5.13).

Certainly $I(A) = I(\limsup \bar{X}_n > \beta$ and $\liminf \bar{X}_n < \alpha)$ is a function of the entire sequence $X_0, X_1, \ldots$. Recall that the limits superior and inferior of a real sequence $a_1, a_2, \ldots$ coincide with the corresponding limits on any shifted sequence $a_{k+1}, a_{k+2}, \ldots$. It follows that

$$I(A) = I(\limsup \bar{X}_{k+n} > \beta \text{ and } \liminf \bar{X}_{k+n} < \alpha), \tag{5.14}$$

for any fixed k, and $I(A)$ is invariant under time shifts in this sense. This will be important later.

Introduce the notation $Y_i = X_i - \beta$, $i = 0, 1, \ldots$, and

$$S_{i,k} = Y_i + Y_{i+1} + \cdots + Y_{i+k-1}, \qquad k \geq 1.$$

There are k summands in $S_{i,k}$, the initial one being Y_i. Let

$$M_{i,n} = \max\{0, S_{i,1}, \ldots, S_{i,n}\}.$$

Of course,

$$\begin{aligned} 0 < S_{0,n} &= Y_0 + \cdots + Y_{n-1} \\ &= (X_0 + \cdots + X_{n-1}) - n\beta, \end{aligned}$$

when and only when

$$\bar{X}_n = n^{-1}(X_0 + \cdots + X_{n-1}) > \beta.$$

The event A entails that the inequality $\bar{X}_n > \beta$ takes place for some positive integer, which in turn implies that the event $\{M_{0,n} > 0\}$ must occur for n sufficiently large. Consequently, we have

$$A \subset \bigcup_{n=1}^{\infty} \{M_{0,n} > 0\},$$

and in view of the monotone nature of $M_{0,n}$

$$I(A) = \lim_{n \to \infty} I(A) I\{M_{0,n} > 0\},$$

where, as usual, $I\{M_{0,n} > 0\}$ denotes the indicator random variable of the event $\{M_{0,n} > 0\}$. Finally, by virtue of the hypothesis $E[|Y_0|] < \infty$, we are permitted to interchange limit and expectation and thereby achieve the formula

$$E[Y_0 I(A)] = \lim_{n \to \infty} E[Y_0 I\{M_{0,m} > 0\} I(A)].$$

We can now proceed to the verification of the two inequalities (5.12) and (5.13).

Obviously,

$$\begin{aligned} M_{1,n} &= \max\{0, S_{1,1}, \ldots, S_{1,n}\} \\ &\geq S_{1,k}, \qquad \text{for} \quad k = 1, \ldots, n-1, \end{aligned}$$

and so

$$Y_0 + M_{1,n} \geq Y_0 + S_{1,k} = S_{0,k+1}, \qquad k = 1, \ldots, n-1.$$

We write this as

$$Y_0 \geq S_{0,k} - M_{1,n}, \qquad \text{for} \quad k = 2, \ldots, n.$$

Trivially, since $M_{1,n} \geq 0$,

$$Y_0 \geq Y_0 - M_{1,n},$$

which, combined with the previous inequality, gives

$$Y_0 \geq \max\{S_{0,1}, \ldots, S_{0,n}\} - M_{1,n},$$

and, since $M_{0,n} = \max\{0, S_{0,1}, \ldots, S_{0,n}\}$,

$$Y_0 \geq M_{0,n} - M_{1,n}, \qquad \text{if} \quad M_{0,n} > 0.$$

Thus

$$\begin{aligned} E[Y_0 I\{M_{0,n} > 0\} I(A)] &\geq E[(M_{0,n} - M_{1,n}) I\{M_{0,n} > 0\} I(A)] \\ &= E[M_{0,n} I(A)] - E[M_{1,n} I\{M_{0,n} > 0\} I(A)] \\ &\geq E[M_{0,n} I(A)] - E[M_{1,n} I(A)], \end{aligned}$$

where the last inequality is a consequence of $M_{1,n} \geq 0$. Recall that the event A is invariant under shifts in time [see (5.14)]. Also since the process is stationary, $M_{0,n}$ and $M_{1,n}$ share the same distribution. Thus $E[M_{0,n} I(A)] = E[M_{1,n} I(A)]$, and the last right-hand side in the above string of inequalities is zero. Hence

$$E[Y_0 I(A)] = \lim_{n \to \infty} E[Y_0 I\{M_{0,n} > 0\} I(A)] \geq 0.$$

Since $Y_0 = X_0 - \beta$, we have demonstrated (5.12), that $E[(X_0 - \beta) I(A)] \geq 0$. The proof of (5.13) is accomplished by applying exactly parallel reasoning to $\tilde{Y}_k = \alpha - X_k$. As these inequalities were all that was needed, the proof of the theorem is complete. ■

Remark 5.1. Under the conditions of the strong ergodic theorem, it is true in addition that

$$\lim_{n \to \infty} E[|\bar{X}_n - \bar{X}|] = 0. \tag{5.15}$$

Hence

$$E[\bar{X}] = E[X_n] = m. \tag{5.16}$$

Remark 5.2. Further elementary considerations show that

$$\bar{X} = \lim_{n \to \infty} \frac{1}{n} (X_k + \cdots + X_{k+n-1}), \tag{5.17}$$

for every $k = 1, 2, \ldots$. Suppose the random variables $X_0, X_1, \ldots$ are, in addition, independent. In view of (5.17), $\bar{X}$ is then independent of $X_0, \ldots, X_{k-1}$ for every k, and consequently independent of $\bar{X}_k$. Since $\bar{X} = \lim \bar{X}_k$, we must have $\bar{X}$ independent of itself. The only possibility, then, is for $\bar{X}$ to be constant (why?), and (5.16) tells us that constant must be m. Thus $\lim_{n \to \infty} \bar{X}_n = m$ for every sequence of independent identically distributed random variables having finite mean m. We have thus verified the strong law of large numbers.

The strong law motivates a desire to find general conditions under which the limit random variable $\bar{X}$ is constant. In the mean square ergodic theorem, this was the result if and only if $\bar{X}_n$ and X_1 were asymptotically uncorrelated. The situation is not this simple under the weaker assumption of the strong ergodic theorem. Nonetheless, some principles can be laid down.

If $x = (x_0, x_1, \ldots)$ is a real sequence, let Tx denote the shifted sequence $(x_1, x_2, \ldots)$. We call T the *shift operator*. A set A of real sequences is called

shift invariant when Tx is an element of A if and only if x is in A. Several examples will kindle the imagination:

$$A_1 = \{x : \text{for some } k = 1, 2, \ldots, x_k = x_{k+1} = \cdots = 0\},$$
$$A_2 = \{x : \limsup x_n = a\},$$
$$A_3 = \{x : \lim n^{-1}(x_1 + \cdots + x_n) = b\}.$$

The student should verify that A_1, A_2 and A_3 are shift invariant sets.

A stationary process is said to be *ergodic* if $\Pr\{(X_0, X_1 \ldots) \in A\}$ is either zero or one whenever A is shift invariant.

Theorem 5.5. *Let $\{X_n\}$ be an ergodic stationary process having a finite mean m. Then, with probability one,*

$$\lim_{n\to\infty} \frac{1}{n}(X_1 + \cdots + X_n) = m,$$

Proof. For each real a define

$$A = \{x = (x_0, x_1 \ldots); \lim_{n\to\infty} n^{-1}(x_1 + \cdots + x_n) \leq a\}.$$

Then A is shift invariant. It follows that

$$\Pr\{(X_0, X_1, \ldots) \in A\} = \Pr\{\lim n^{-1}(X_1 + \cdots + X_n) \leq a\}$$
$$= \Pr\{\bar{X} \leq a\} = 0 \quad \text{or} \quad 1,$$

for every constant a. Hence $\bar{X}$ is necessarily a constant random variable. In view of (5.16), that constant can only be m. ∎

It is obviously desirable to have some equivalent and more accessible formulations of ergodicity. Unfortunately not much can be done here. We state some results, without proof, in the next theorem.

Theorem 5.6. *Let $\{X_n\}$ be a stationary process. The following conditions are equivalent:*

(a) *$\{X_n\}$ is ergodic;*

(b) *For every shift invariant set A,*

$$\Pr\{(X_0, X_1, \ldots) \in A\} = 0 \quad \text{or} \quad 1;$$

(c) *For every set A of real sequences $(x_0, x_1, \ldots)$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I\{(X_j, X_{j+1}, \ldots) \in A\} = \Pr\{(X_0, X_1, \ldots) \in A\};$$

(d) *For every* $k = 1, 2, \ldots$ *and every set* A *of real vectors* $(x_0, \ldots, x_k)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I\{(X_j, \ldots, X_{j+k}) \in A\} = \Pr\{(X_0, \ldots, X_k) \in A\};$$

(e) *For every* k *and every function* ϕ *of* $k+1$ *variables,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \phi(X_j, \ldots, X_{j+k}) = E[\phi(X_0, \ldots, X_k)],$$

provided the expectation exists;

(f) *For every function* ϕ *of a real sequence* $(x_0, x_1, \ldots)$.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \phi(X_j, X_{j+1}, \ldots) = E[\phi(X_0, X_1, \ldots)],$$

provided the expectation exists.

In these conditions, the existence of the limits is a consequence of the strong ergodic theorem. For example, if ϕ is a function of $k+1$ variables for which $E[|\phi(X_0, \ldots, X_k)|] < \infty$, then

$$Y_n = \phi(X_n, \ldots, X_{n+k}), \qquad n = 0, 1, \ldots,$$

determines a stationary process to which the strong ergodic theorem applies. Hence, $\lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} Y_j$ exists. What is asserted in condition (e) of Theorem 5.6 is that this limit is constant.

Remark 5.3. It is obvious from the equivalent condition of Theorem 5.6 that, for any function ϕ, the sequence

$$Y_n = \phi(X_n, X_{n+1}, \ldots)$$

generates an ergodic stationary process whenever $\{X_n\}$ is ergodic and stationary.

A stationary process $\{X_n\}$ is said to be *mixing* (or *strong mixing*) if for all sets A and B of real sequences

$$\lim_{n\to\infty} \Pr\{(X_1, X_2, \ldots) \in A \text{ and } (X_{n+1}, X_{n+2}, \ldots) \in B\}$$
$$= \Pr\{(X_1, X_2, \ldots) \in A\} \Pr\{(X_1, X_2, \ldots) \in B\}.$$

Mixing is a form of asymptotic independence, and therefore that every mixing process is ergodic is perhaps not surprising. We now validate this

statement. Suppose $A = B$ is an invariant set. Then $(X_n, X_{n+1}, \ldots) \in B$ if and only if $(X_1, X_2, \ldots) \in B$. The mixing property applies, yielding

$$\Pr\{(X_1, X_2, \ldots) \in A\} = \Pr\{(X_1, X_2, \ldots) \in A, (X_n, X_{n+1}, \ldots) \in A\}$$
$$\to [\Pr\{(X_1, X_2 \ldots) \in A\}]^2, \qquad \text{as} \quad n \to \infty.$$

The only possibilities are $\Pr\{(X_1, X_2, \ldots) \in A\} = 0$ or 1, so the process is ergodic. There are other equivalent formulations for mixing. For example, it is enough to check that, for arbitrary $k = 1, 2, \ldots$ and sets A, B in k-space, that

$$\lim_{n \to \infty} \Pr\{(X_1, \ldots, X_k) \in A \text{ and } (X_{n+1}, \ldots, X_{n+k}) \in B\}$$
$$= \Pr\{(X_1, \ldots, X_k) \in A\} \Pr\{(X_1, \ldots, X_k) \in B\}.$$

We would be remiss in our duties if we did not at least indicate that what we have exposed heretofore is not ergodic theory, but mostly the application of ergodic theory to stationary processes. The modern view is much more general. Let L be an abstract space of functions f for which some notion of integral $\int f$ is defined. Let T be an operator that carries a function f in L into the function Tf in L. Ergodic theory is the study of iterates T^n of the operators T that satisfy:

(i) If f is a nonnegative function then so is Tf, and

(ii) $\int |Tf| \leq \int |f|$.

In the context of a stationary process $\{X_n\}$, L consists of all functions $f(\{X_n\}) = \phi(X_0, X_1, \ldots)$ for which $E[|\phi(X_0, X_1, \ldots)|] < \infty$, the notion of integral is the expectation, $\int f(\{X_n\}) = E[\phi(X_0, X_1, \ldots)]$, and T is related to the shift operator by

$$T\phi(X_0, X_1, X_2, \ldots) = \phi(X_1, X_2, \ldots).$$

6: Applications of Ergodic Theory

At the start of Section 5 mention was made of the ergodic theorem used to justify time averages such as $\bar{X}_n = n^{-1}(X_0 + \cdots + X_{n-1})$ as estimators of the corresponding process expectations, in this case $m = E[X_0]$. There are numerous other applications. We develop three examples.

A. BRANCHING PROCESSES IN RANDOM ENVIRONMENTS

A key feature of the branching processes introduced in Example F of Chapter 2 and thoroughly discussed in Chapter 8 is the invariance over time of the offspring distribution. But there are a multitude of situations

in which the offspring distribution depends on a changing environment. As an introduction to recent work in this area, we treat the extinction problem in an environment that varies randomly according to a stationary ergodic process.

We postulate a stationary ergodic process $\bar{\zeta} = \{\zeta_n\}_{n=0}^{\infty}$, called the *environmental process*. To each value ζ_n there corresponds an offspring distribution $p_{\zeta_n} = \{p_n(i)\}_{i=0}^{\infty}$. Of course, $p_n(i) \geq 0$ and $\sum_{i=0}^{\infty} p_n(i) = 1$. Then $\{p_{\zeta_n}\}_{n=0}^{\infty}$ is again a stationary ergodic process, albeit one whose values are discrete probability distributions.

Let $Z_n (= Z(n))$ count the number of particles existing in the nth generation, and take $Z_0 = 1$. Conditional on a prescribed environment $\bar{\zeta} = \{\zeta_n\}$, we postulate that $\{Z_n\}$ evolves as an ordinary branching process excepting only that the offspring distribution in generation n is p_{ζ_n}. To be precise, we assume

$$E[s^{Z(n+1)} | \zeta_0, \ldots, \zeta_n; Z_0, \ldots, Z_n] = [\phi_{\zeta_n}(s)]^{Z(n)}, \qquad n = 0, 1, 2, \ldots, |s| \leq 1,$$

where

$$\phi_{\zeta_n}(s) = \sum_{j=0}^{\infty} p_n(j) s^j$$

is the probability generating function corresponding to p_{ζ_n}. In other words, conditioned on the past environment and population levels, $Z(n+1)$ is the sum of $Z(n)$ litters (or broods) of offspring, where the sizes of each brood are independent and identically distributed random variables, following the common distribution p_{ζ_n}.

Let B_n be the event that $Z_n = 0$, and clearly $B = \bigcup_{n=1}^{\infty} B_n$ connotes the event of (ultimate) extinction. It is useful to introduce the notation

$$q(\bar{\zeta}) = \Pr\{B | \bar{\zeta}\}$$

and

$$q = \Pr\{B\} = E[q(\bar{\zeta})].$$

Conditioned on the environmental process $\bar{\zeta}$, $\{Z_n\}$ behaves as an ordinary, although *time-inhomogeneous*, branching process, and the techniques of Chapter 8 lead us easily to the following formula*:

$$E[s^{Z(n+1)} | \bar{\zeta}] = [\phi_{\zeta_0}(\phi_{\zeta_1}(\ldots (\phi_{\zeta_n}(s)) \ldots))].$$

* We assume throughout that for any Z, the associated $\phi_{\zeta_i}(s)$ are never constant functions.

We frequently write ϕ_i for brevity in place of ϕ_{ζ_i} if no ambiguity is likely in the interpretation. Obviously

$$q(\bar{\zeta}) = \lim_{n\to\infty} \phi_0(\phi_1(\dots(\phi_n(0))\dots)) = \phi_{\zeta_0}\Big[\lim_{n\to\infty} \phi_1\phi_2(\dots(\phi_n(0)))\dots)\Big],$$

or

$$q(\bar{\zeta}) = \phi_{\zeta_0}[q(T\bar{\zeta})], \tag{6.1}$$

where $T\bar{\zeta}$ denotes the shifted environmental sequence $\{\zeta_1, \zeta_2, \dots\}$. The important functional equation $q(\bar{\zeta}) = \phi_{\zeta_0}[q(T\bar{\zeta})]$ is the stochastic analog of the equation $s = \phi(s)$ of Chapter 8.

Consider the event $\{q(\bar{\zeta}) = 1\}$. In words, this event occurs whenever the random environment leads to sure (i.e., certain) extinction of the population. We may clearly exploit the properties of the probability generating function ϕ_{ζ_0} in the basic equation $q(\bar{\zeta}) = \phi_{\zeta_0}[q(T\bar{\zeta})]$ to conclude that $q(T\bar{\zeta}) = 1$ whenever $q(\bar{\zeta}) = 1$. That is, the event $\{q(\bar{\zeta}) = 1\}$ is shift invariant. By assumption, the environmental process $\bar{\zeta} = \{\zeta_n\}$ is ergodic, and therefore we have

$$\Pr\{q(\bar{\zeta}) = 1\} = 0 \quad \text{or} \quad 1. \tag{6.2}$$

Let $m_{\zeta_n} = \sum_{j=0}^{\infty} jp_n(j)$ be the conditional mean, given the environment, of the nth generation's offspring distribution. The $\{m_{\zeta_n}\}$ is again a stationary ergodic process, and if

$$E[|\log m_{\zeta_0}|] < \infty, \tag{6.3}$$

the strong ergodic theorem tells us

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \log m_{\zeta_k} = E[\log m_{\zeta_0}].$$

Let us assume (6.3) and, in addition,

$$\Pr\{q(\bar{\zeta}) < 1\} = 1, \tag{6.4}$$

and see where this leads us. Probability generating functions are increasing and convex so that

$$m_{\zeta_0} = \phi'_{\zeta_0}(1) \geq \frac{1 - \phi_{\zeta_0}(s)}{1 - s}, \qquad \text{for any} \quad s \in [0, 1). \tag{6.5}$$

According to (6.4), $q(\bar{\zeta}) < 1$, and added to the basic functional equation $q(\bar{\zeta}) = \phi_{\zeta_0}[q(T\bar{\zeta})]$, we stipulated equivalent to (6.4) that $q(T\bar{\zeta}) < 1$ (see

the footnote on p. 490). This allows the substitution $q(\bar{\zeta}) = q(T\bar{\zeta})$ from (6.1) in (6.5) to obtain

$$m_{\zeta_0} \geq \frac{1 - \phi_{\zeta_0}[q(T\bar{\zeta})]}{1 - q(T\bar{\zeta})} = \frac{1 - q(\bar{\zeta})}{1 - q(T\bar{\zeta})}.$$

The stationary property extends this to

$$m_{\zeta_n} \geq \frac{1 - q(T^n\bar{\zeta})}{1 - q(T^{n+1}\bar{\zeta})},$$

where $T^n\bar{\zeta}$ is the sequence $(\zeta_n, \zeta_{n+1}, \ldots)$. A collapsing sum appears in

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{n-1} \log m_{\zeta_k} &\geq \frac{1}{n}\sum_{k=0}^{n-1} \log\left\{\frac{1 - q(T^k\bar{\zeta})}{1 - q(T^{k+1}\bar{\zeta})}\right\} \\ &= \frac{1}{n}[\log(1 - q(\bar{\zeta})) - \log(1 - q(T^n\bar{\zeta}))] \\ &\geq \frac{1}{n}\log[1 - q(\bar{\zeta})].\end{aligned}$$

If $E[|\log m_{\zeta_0}|] < \infty$, as we have assumed, then the left-hand side converges to $E[\log m_{\zeta_0}]$ by the strong ergodic theorem, while the right vanishes. Thus

$$E[\log m_{\zeta_0}] \geq 0.$$

It is possible to strengthen the conclusion by ruling out the possibility $E[\log m_{\zeta_0}] = 0$. To avoid a more technical proof, let us assume

$$E[|\log[1 - q(\bar{\zeta})] - \log[1 - q(T\bar{\zeta})]|] < \infty.$$

Then

$$Y_n = \log[1 - q(T^n\bar{\zeta})] - \log[1 - q(T^{n+1}\bar{\zeta})]$$

forms an ergodic stationary process to which the ergodic theorem applies. Thus

$$\begin{aligned}E\left[\log\frac{1 - q(\bar{\zeta})}{1 - q(T\bar{\zeta})}\right] &= \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\log\left(\frac{1 - q(T^k\bar{\zeta})}{1 - q(T^{k+1}\bar{\zeta})}\right) \\ &= \lim_{n\to\infty}\frac{1}{n}\{\log[1 - q(\bar{\zeta})] - \log[1 - q(T^n\bar{\zeta})]\} \qquad (6.6) \\ &\geq 0.\end{aligned}$$

Now from (6.5) we know that

$$\log \phi'_{\zeta_0}(1) - \log\left\{\frac{1 - \phi_{\zeta_0}[q(T\bar{\zeta})]}{1 - q(T\bar{\zeta})}\right\} \geq 0.$$

If we assume $0 = E[\log m_{\zeta_0}] = E[\log \phi'_{\zeta_0}(1)]$, then the inequality

$$0 \leq -E\left[\log\left\{\frac{1-\phi_{\zeta_0}[q(T\bar{\zeta})]}{1-q(T\bar{\zeta})}\right\}\right]$$

ensues. Comparing to (6.6), we conclude that

$$E\left[\log\left\{\frac{1-\phi_{\zeta_0}[q(T\bar{\zeta})]}{1-q(T\bar{\zeta})}\right\}\right] = 0,$$

and consequently the only consistent value is $q(T\bar{\zeta}) = 1$ in violation of the stipulation $\Pr\{q(\bar{\zeta}) < 1\} = 1$. Therefore, $E[\log m_{\zeta_0}] > 0$ must hold.

To sum up, if $E[\log m_{\zeta_0}] < \infty$, then $\Pr\{q(\bar{\zeta}) < 1\} = 1$ implies $E[\log m_{\zeta_0}] > 0$. Conversely, $E[\log m_{\zeta_0}] \leq 0$, entails $\Pr\{q(\bar{\zeta}) = 1\} = 1$. [Use (6.2).]

B. THE RANGE OF A RANDOM WALK

Let $X_1, X_2, \ldots$ be independent identically distributed random variables whose possible values are the integers $0, \pm 1, \pm 2, \ldots$. Consider the partial sum process $S_n = X_1 + \cdots + X_n$, $n \geq 1$, and $S_0 = 0$. Define the range R_n involving the first n sums to be the number of distinct values in $\{S_1, \ldots, S_n\}$. We may write

$$R_n = \sum_{k=1}^{n} I_k$$

where

$$I_k = \begin{cases} 1, & \text{if } S_j \neq S_k, \text{ for } j = 1, \ldots, k-1, \\ 0, & \text{otherwise.} \end{cases}$$

Now

$$\begin{aligned} E[I_k] &= \Pr\{S_k - S_{k-1} \neq 0, S_k - S_{k-2} \neq 0, \ldots, S_k - S_1 \neq 0\} \\ &= \Pr\{X_k \neq 0, X_k + X_{k-1} \neq 0, \ldots, X_k + \cdots + X_2 \neq 0\} \\ &= \Pr\{S_1 \neq 0, S_2 \neq 0, \ldots, S_{k-1} \neq 0\} \end{aligned}$$

(since the X_i are independent identically distributed). We obtain

$$\lim_{k\to\infty} E[I_k] = \Pr\{S_n \neq 0 \text{ for all } n \geq 1\}. \tag{6.7}$$

Observe that $\{S_n \neq 0 \text{ for all } n \geq 1\}$ is the event that the $\{S_n\}$ process never returns to its origin $S_0 = 0$. But every convergent sequence also converges in the Cesaro mean sense. From (6.7) then,

$$\lim_{n\to\infty} \frac{E[R_n]}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E[I_k] = \Pr\{S_n \neq 0 \text{ for all } n \geq 1\}. \tag{6.8}$$

It is a deeper fact that this limit holds even if expectations are not taken on the left. That is, with probability one,

$$\lim_{n\to\infty}\frac{R_n}{n}=\Pr\{S_n\neq 0 \text{ for all } n\geq 1\}. \tag{6.9}$$

We verify this by proving the equivalent pair of inequalities for the limit superior and limit inferior. First, designate m as any positive integer and let Z_k be the number of distinct points visited by the sequence of sums $\{S_i\}$ during the times $(k-1)m+1$ to km. Equivalently, Z_k is the range involved in the collection $\{S_{(k-1)m+1}, \ldots, S_{km}\}$. Note that Z_k depends only on X_n for n between $(k-1)m+1$ and km, so that the Z_k are mutually independent random variables, $|Z_k|\leq m$, and, manifestly, the Z_k are identically distributed. Verify the inequality $R_{km}\leq Z_1+\cdots+Z_k$ and apply the law of large numbers to obtain

$$\limsup_{k\to\infty}\frac{R_{km}}{km}\leq\limsup_{k\to\infty}\frac{Z_1+\cdots+Z_k}{mk}=\frac{E[Z_1]}{m}.$$

With $[n/m]$ denoting the largest integer not exceeding n/m, we have $R_n\leq R_{([n/m]+1)m}$, whence

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}R_n&\leq\limsup_{n\to\infty}\frac{([n/m]+1)m}{n}\,\frac{R_{([n/m]+1)m}}{([n/m]+1)m}\\&=\frac{E[Z_1]}{m}.\end{aligned}$$

This holds for any m. But $Z_1=R_m$, so by (6.8)

$$\limsup_{n\to\infty}\frac{1}{n}R_n\leq\limsup_{n\to\infty}\frac{E[R_n]}{n}=\Pr\{S_n\neq 0 \text{ for all } n\geq 1\}. \tag{6.10}$$

The reverse inequality is the deeper one, and for it we invoke the strong ergodic theorem.

Define

$$V_k=\begin{cases}1, & \text{if } S_j\neq S_k, \text{ for all } j>k,\\ 0, & \text{otherwise.}\end{cases}$$

V_k is one, if S_k is a state never visited after time k, and zero, otherwise, and $V_1+\cdots+V_n$ is the number of states visited up to time n that are never revisited. Since R_n is the number of states visited in time n that are not revisited prior to $n+1$, manifestly $R_n\geq V_1+\cdots+V_n$. Now

$\{X_k\}$ is a stationary ergodic process (see Problem 21) and by Remark 5.3 so is $\{V_k\}$, since

$$V_k = \begin{cases} 1, & \text{if } X_{k+1} \neq 0, \quad X_{k+1} + X_{k+2} \neq 0, \ldots, \\ 0, & \text{otherwise,} \end{cases}$$
$$= \phi(X_{k+1}, X_{k+2}, \ldots),$$

where

$$\phi(x_1, x_2, \ldots) = \begin{cases} 1, & \text{if } x_1 \neq 0, \quad x_1, + x_2 \neq 0, \quad x_1 + x_2 + x_3 \neq 0, \cdots \\ 0, & \text{otherwise.} \end{cases}$$

The ergodic theorem implies

$$\liminf_{n\to\infty} \frac{R_n}{n} \geq \liminf_{n} \frac{V_1 + \cdots + V_n}{n} = E[V_1]. \tag{6.11}$$

Of course,

$$E[V_1] = \Pr\{S_k \neq 0, \text{ for all } k \geq 1\}.$$

Combining (6.10) and (6.11), we see that the proof of (6.9) is done.

C. ENTROPY

Probability measures uncertainty about the *occurrence* of a single event. Entropy measures the uncertainty of a collection of events. Let X be a random variable assuming the value i with probability p_i, $i = 1, \ldots, n$. The *entropy of* X, (or of the events $\{X = i\}$ $i = 1, \ldots, n$) is computed according to

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{6.12}$$

(with the understanding that $0 \cdot \log 0 = 0$). The reader should verify that the definition possesses the following three desirable properties: (i) The entropy of a constant random variable is zero; (ii) If we add the possible value $i + 1$, assigning it probability $p_{i+1} = 0$, the entropy is unchanged; and (iii) The entropy is maximized, with maximum value $\log n$, when $p_1 = \cdots = p_n = 1/n$. This last property agrees with our intuition about uncertainty, e.g., the random variable

$$X_1 = \begin{cases} 1, & \text{with probability } 0.999, \\ 0, & \text{with probability } 0.001, \end{cases}$$

is far more predictable or ascertainable than

$$X_2 = \begin{cases} 1, & \text{with probability } \frac{1}{2} \\ 0, & \text{with probability } \frac{1}{2} \end{cases}$$

It is natural to extend this definition to a pair of random variables X, Y, where $\Pr\{X=i,\ Y=j\}=p_{ij}$, through the formula

$$H(X,\ Y)=-\sum_{i,\ j}p_{ij}\log p_{ij}.$$

Define the *conditional entropy of X given Y* by

$$H(X|\,Y)=-\sum_{j}\Pr\{Y=j\}\sum_{i}p(i|\,j)\log p(i|\,j),$$

where $p(i|\,j)=\Pr\{X=i|\,Y=j\}$. Substituting $p(i|\,j)=p_{ij}/\Pr\{Y=j\}$ gives

$$\begin{aligned}H(X|\,Y)&=-\sum_{i,\ j}p_{ij}\log[p_{ij}/\Pr\{Y=j\}]\\&=H(X,\ Y)-H(Y).\end{aligned}\tag{6.13}$$

What is important here is that

$$H(X,\ Y)\geq H(Y).\tag{6.14}$$

In words, the uncertainty in a pair of random variables exceeds that of either one.

It is a little harder to show that X given Y has less uncertainty than X unconditionally. Let $q_j=\Pr\{Y=j\}$. Observe that the function $\Gamma(t)=-t\log t$ is concave for $t>0$, and use Jensen's inequality (i.e., the extended definition of convexity, see page 249) to obtain

$$\begin{aligned}H(X|\,Y)&=\sum_{i}\sum_{j}q_j\Gamma[p(i|\,j)]\\&\leq\sum_{i}\Gamma\left[\sum_{j}q_jp(i|\,j)\right]\\&=H(X).\end{aligned}\tag{6.15}$$

It follows from this and (6.13) that

$$H(X,\ Y)\leq H(X)+H(Y).\tag{6.16}$$

We continue to extend the definition of entropy. If $X_1,\ \ldots,\ X_m$ are jointly distributed random variables and $\Pr\{X_1=i_1,\ \ldots,\ X_m=i_m\}=p(i_1,\ \ldots,\ i_m)$, set

$$H(X_1,\ \ldots,\ X_m)=-\sum_{i_1,\ldots,i_m}p(i_1,\ \ldots,\ i_m)\log p(i_1,\ \ldots,\ i_m).$$

We have the results analogous to (6.13) and (6.15),

$$\begin{aligned}&H(X_1,\ \ldots,\ X_m)\\&=H(X_1)+H(X_2|X_1)+H(X_3|X_1,\ X_2)+\cdots+H(X_m|X_1,\ \ldots,\ X_{m-1}),\end{aligned}\tag{6.17}$$

and

$$H(X_k|X_1,\ \ldots,\ X_{k-1})\leq H(X_k|X_2,\ \ldots,\ X_{k-1}).\tag{6.18}$$

Now let $\{X_n\}$ be a stationary process where the possible values of each X_k are the numbers $1, \ldots, N$. The stationary property in conjunction with (6.18) implies

$$H(X_{k-1}|X_1, \ldots, X_{k-2}) = H(X_k|X_2, \ldots, X_{k-1}) \geq H(X_k|X_1, \ldots, X_{k-1}).$$

Since the sequence is monotone decreasing and nonnegative, we may define the entropy of the process $\{X_n\}$ by

$$H(\{X_n\}) = \lim_{k\to\infty} H(X_k|X_1, \ldots, X_{k-1}). \tag{6.19}$$

Alternatively, we may write

$$H(\{X_n\}) = \lim_{l\to\infty} \frac{1}{l} H(X_1, \ldots, X_l),$$

since according to (6.17), $(1/l)H(X_1, \ldots, X_l)$ is the Cesaro average of $H(X_k|X_1, \ldots, X_{k-1})$.

Let $p(i_1, \ldots, i_m) = \Pr\{X_1 = i_1, \ldots, X_m = i_m\}$. The remarkable result we have been leading to is, with probability one,

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log p(X_0, \ldots, X_{n-1})\right] = H(\{X_n\}), \tag{6.20}$$

provided $\{X_n\}$ is ergodic. The right member of (6.20) is constant, computed according to (6.19). The left is a limit taken along a random sequence $X_1, X_2, \ldots$. The result says that, for large n, the probability $p(X_1, \ldots, X_n)$ of the observed sequence $X_1, X_2, \ldots$ is bound to be near $\exp(-nH(\{X_n\}))$.

We continue this example by verifying (6.20) in the special case that $\{X_n\}$ is a stationary ergodic finite Markov chain. Later we will state and prove the general result as a theorem.

Suppose $P = \|P(i, j)\|_{i,j=1}^N$ is the transition matrix of an irreducible finite state Markov chain $\{X_n\}$. We assume $\Pr\{X_0 = i\} = \pi(i)$, where $\{\pi(i)\}_1^N$ is the stationary distribution associated with P. Compute

$$H(X_n|X_0, \ldots, X_{n-1}) = -\sum \pi(i_0)P(i_0, i_1) \cdots P(i_{n-2}, i_{n-1}) \sum_{i_n} P(i_{n-1}, i_n) \log P(i_{n-1}, i_n)$$

(by virtue of the Markov chain character of the process and using the fact that $\pi(i)$ is a stationary distribution)

$$= -\sum_{i,j} \pi(i)P(i, j) \log P(i, j),$$

and so

$$H(\{X_n\}) = -\sum_{i,j} \pi(i)P(i, j) \log P(i, j).$$

On the other hand,

$$-\frac{1}{n}\log p(X_0,\ldots,X_{n-1}) = -\frac{1}{n}\log\{\pi(X_0)P(X_0,X_1)\cdots P(X_{n-2},X_{n-1})\}$$

$$= \frac{1}{n}\sum_{i=0}^{n-2} W_i - \frac{1}{n}\log\pi(X_0),$$

where

$$W_i = -\log P(X_i, X_{i+1}).$$

An irreducible finite-state Markov chain started with its stationary distribution generates an ergodic stationary process. Since there are only a finite number of states, W_i is bounded. The ergodic theorem applies to yield

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log p(X_0,\ldots,X_{n-1})\right]$$

$$= \lim_{n\to\infty}\left[-\frac{1}{n}\sum_{k=0}^{n-2} W_k\right]$$

$$= E[W_0] = -\sum_{i,j}\pi(i)P(i,j)\log P(i,j)$$

$$= H(\{X_n\}),$$

as desired. Here is the general case.

Theorem 6.1. *Let $\{X_n\}$ be an ergodic stationary process having a finite state space $\{1,\ldots,N\}$. Let $p(i_1,\ldots,i_m) = \Pr\{X_1 = i_1,\ldots,X_m = i_m\}$, and*

$$H(\{X_n\}) = \lim_{n\to\infty}\left(-\frac{1}{n}\sum_{k=1}^{n}\sum_{i_1,\ldots,i_k} p(i_1,\ldots,i_k)\log p(i_1,\ldots,i_k)\right).$$

Then, with probability one,

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log p(X_1,\ldots,X_n)\right] = H(\{X_n\}).$$

Proof. Note first, imposing no limitations on the analysis, we may as well assume $n = \ldots,-1,0,+1,\ldots$. Indeed, given any one-sided stationary process $\{\tilde{X}_n;\ n = 0, 1, \ldots\}$, a consistent collection of finite-dimensional distributions for a two-sided stationary process is given by

$$\Pr\{X_m = i_m,\ldots,X_{m+k} = i_{m+k}\} = \Pr\{\tilde{X}_1 = i_m,\ldots,\tilde{X}_{k+1} = i_{m+k}\},$$

for any $m = \ldots,-1,0,+1,\ldots$ and $k = 1, 2, \ldots$.

Set

$$f_k(i; i_1, \ldots, i_k) = \Pr\{X_k = i | X_0 = i_k, \ldots, X_{k-1} = i_1\}$$
$$= \Pr\{X_0 = i | X_{-k} = i_k, \ldots, X_{-1} = i_1\}$$

(by stationarity). The backward martingale convergence theorem (Theorem 8.1 of Chapter 6), assures us of the existence of

$$f(i; i_1, i_2, \ldots) = \lim_{k \to \infty} f_k(i; i_1, \ldots, i_k)$$
$$= \Pr\{X_0 = i | X_{-1} = i_1, X_{-2} = i_2, \ldots\}$$
$$= \Pr\{X_n = i | X_{n-1} = i_1, X_{n-2} = i_2, \ldots\}.$$

Set

$$g_0(i) = -\log \Pr\{X_0 = i\},$$
$$g_k(i_0, i_1, \ldots, i_k) = -\log f_k(i_0; i_1, \ldots, i_k)$$
$$= -\log \Pr\{X_0 = i_0 | X_{-k} = i_k, \ldots, X_{-1} = i_1\}$$
$$= -\log \frac{\Pr\{X_{-k} = i_k, \ldots, X_{-1} = i_1, X_0 = i_0\}}{\Pr\{X_{-k} = i_k, \ldots, X_{-1} = i_1\}},$$

and

$$g(i, i_1, i_2, \ldots) = -\log f(i; i_1, i_2, \ldots).$$

Direct computation produces the equation

$$-\frac{1}{n} \log p(X_0, \ldots, X_{n-1}) = \frac{1}{n} \sum_{k=0}^{n-1} g_k(X_k, \ldots, X_0).$$

If we could replace $g_k(X_k, \ldots, X_0)$ by $g(X_k, X_{k-1}, \ldots)$, the right member would be the type of average to which the ergodic theorem applies, $W_k = g(X_k, X_{k-1}, \ldots)$ being stationary since $\{X_k\}$ is stationary. This adjustment was immediate in the Markov case since there ipso facto $g_k(X_k, \ldots, X_0) = g(X_k, X_{k-1}, \ldots)$ whenever $k \geq 1$.

In our present situation our first task will be to establish

$$E\left[\sup_{n \geq 1} g_n(X_n, \ldots, X_0)\right] < \infty. \tag{6.21}$$

Fix $\lambda > 0$ and decompose the event as indicated:

$$\left\{\sup_{n \geq 1} g_n(X_n, \ldots, X_0) > \lambda\right\} = \bigcup_{n=1}^{\infty} \{g_k(X_k, \ldots, X_0) \leq \lambda \text{ for } k = 1, \ldots, n-1$$
$$\text{and } g_n(X_n, \ldots, X_0) > \lambda\}.$$
$$= \bigcup_{n=1}^{\infty} \{(X_0, \ldots, X_n) \in E_n\},$$

where

$$E_n = \{(i_0, \ldots, i_n); g_k(i_k, \ldots, i_0) \leq \lambda \quad \text{for} \quad k = 1, \ldots, n-1 \quad \text{and} \quad g_n(i_n, \ldots, i_0) > \lambda\}.$$

We have partitioned the event on the left on the basis of the first time n for which $g_n(X_n, \ldots, X_0) > \lambda$. As this is a disjoint union

$$\Pr\left\{\sup_{n \geq 1} g_n(X_n, \ldots, X_0) > \lambda\right\} = \sum_{n=1}^{\infty} \Pr\{(X_0, \ldots, X_n) \in E_n\}.$$

Let us partition further by specifying the final state. Let

$$E_n^{(i)} = \{(i_0, \ldots, i_{n-1}); (i_0, \ldots, i_{n-1}, i) \in E_n\}.$$

Consider now

$$\Pr\{(X_0, \ldots, X_n) \in E_n\} = \sum_i \Pr\{(X_0, \ldots, X_{n-1}) \in E_n^{(i)} \quad \text{and} \quad X_n = i\}. \tag{6.22}$$

By the definition of conditional probability, we have

$$\begin{aligned} &\Pr\{X_n = \mathrm{i} \quad \text{and} \quad (X_0, \ldots, X_{n-1}) \in E_n^{(i)}\} \\ &\quad = \sum_{(i_0, \ldots, i_{n-1}) \in E_n^{(i)}} \Pr\{X_n = i | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} \\ &\quad \cdot \Pr\{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\}. \end{aligned} \tag{6.23}$$

The process is stationary, so the conditional probability may be shifted to read

$$\begin{aligned} &\Pr\{X_n = i | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} \\ &\quad = \Pr\{X_0 = i | X_{-n} = i_0, \ldots, X_{-1} = i_{n-1}\} \\ &\quad = \exp[-g_n(i, i_{n-1}, \ldots, i_0)]. \end{aligned}$$

Manifestly, $g_n(i, i_{n-1}, \ldots, i_0)$ exceeds λ whenever $(i_0, \ldots, i_{n-1})$ is in $E_n^{(i)}$. Thus the probability in (6.23) is bounded by

$$\begin{aligned} &\Pr\{X_n = i \quad \text{and} \quad (X_0, \ldots, X_{n-1}) \in E_n^{(i)}\} \\ &\quad \leq \sum_{(i_0, \ldots, i_{n-1}) \in E_n^{(i)}} e^{-\lambda} \Pr\{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} \\ &\quad \leq e^{-\lambda} \Pr\{(X_0, \ldots, X_{n-1}) \in E_n^{(i)}\}. \end{aligned}$$

For each fixed i, the sets $E_1^{(i)}$, $E_2^{(i)}$, ... are disjoint for the same reasons that E_1, E_2, ... are disjoint. We sum (6.22) over n and use our bound to deduce

$$\begin{aligned}\sum_{n=1}^{\infty} \Pr\{(X_0, \ldots, X_n) \in E_n\} &= \sum_i \sum_{n=1}^{\infty} \Pr\{(X_0, \ldots, X_{n-1}) \in E_n^{(i)} \text{ and } X_n = i\} \\ &\leq e^{-\lambda} \sum_i \left\{ \sum_{n=1}^{\infty} \Pr\{(X_0, \ldots, X_{n-1}) \in E_n^{(i)}\} \right\} \\ &\leq e^{-\lambda} \sum_i 1 \\ &\leq N e^{-\lambda},\end{aligned}$$

where N is the number of distinct states (i values) of the process. We have thus shown

$$\Pr\left\{\sup_{n \geq 1} g_n(X_n, \ldots, X_0) > \lambda\right\} \leq N e^{-\lambda}.$$

With this estimate the validation of (6.21) follows readily.

Since, with probability one,

$$g_k(X_0, X_{-1}, \ldots, X_{-k}) \to g(X_0, X_{-1}, \ldots),$$

as $k \to \infty$, and the expectations of the left functions are uniformly integrable, we may interchange expectation and limit to conclude

$$\begin{aligned}E[g(X_0, X_{-1}, \ldots)] &= \lim_{k \to \infty} E[g_k(X_0, \ldots, X_{-k})] \\ &= \lim_{n \to \infty} \frac{1}{n} E\left\{\sum_{k=0}^{n-1} g_k(X_k, \ldots, X_0)\right\} = H(\{X_k\}).\end{aligned}$$

The second equation is due to stationarity.

Write

$$\begin{aligned}\frac{1}{n} \sum_{k=0}^{n-1} g_k(X_k, \ldots, X_0) = \frac{1}{n} \sum_{k=0}^{n-1} g(X_k, X_{k-1}, \ldots) + \frac{1}{n} \sum_{k=0}^{n-1} \{g_k(X_k, \ldots, X_0) \\ - g(X_k, X_{k-1}, \ldots)\}. \qquad (6.24)\end{aligned}$$

We have established that $E[g(X_k, X_{k-1}, \ldots)] < \infty$. The ergodic theorem implies

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} g(X_k, X_{k-1}, \ldots) = E[g(X_0, X_{-1}, \ldots)] = H(\{X_n\}). \qquad (6.25)$$

Set

$$\phi_N(x_0, x_{-1}, \ldots) = \sup_{k \geq N} |g_k(x_0, \ldots, x_{-k}) - g(x_0, x_{-1}, \ldots)|,$$

and

$$Z_k^N = \phi_N(X_k, X_{k-1}, \ldots).$$

Of course, $\{Z_k^N\}$ is stationary, ergodic, and $E[|Z_k^N|] < \infty$. Then

$$\begin{aligned} \limsup_{n \to \infty} \left| \frac{1}{n} \sum_{k=0}^{n-1} \{g_k(X_k, \ldots, X_0) - g(X_k, X_{k-1}, \ldots)\} \right| \\ \leq \limsup_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} |g_k(X_k, \ldots, X_0) - g(X_k, X_{k-1}, \ldots)| \\ \leq \limsup_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} Z_k^N = E[Z_1^N]. \end{aligned}$$

But as $N \to \infty$, $Z_1^N \to 0$, and the interchange of limit and expectation can be justified to conclude $\lim_{N \to \infty} E[Z_1^N] = 0$. It follows that the second term on the right in (6.24) goes to zero as $n \to \infty$. Combined with (6.25) we have proved

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} g_k(X_k, \ldots, X_0) = H(\{X_n\})$$

as claimed.

7: *Spectral Analysis of Covariance Stationary Processes*

In this section we motivate a canonical representation for zero-mean covariance stationary processes in terms of harmonic functions. Earlier we suggested that an arbitrary such process can be represented as the mean square limit of a sequence of processes of the form of Example B of Section 1, that is, of the form

$$X_n = \sum_{k=0}^{m} \{A_k \cos n\omega_k + B_k \sin n\omega_k\}, \tag{7.1}$$

where $0 \leq \omega_k \leq \pi$, and the coefficients $\{A_k\}$ and $\{B_k\}$ are uncorrelated with zero means and variances $\sigma_k^2 = E[A_k^2] = E[B_k^2]$. In Section 1 we computed the covariance function of (7.1) to be

$$R(v) = \sigma^2 \sum_{k=0}^{m} p_k \cos v\omega_k, \tag{7.2}$$

where $\sigma^2 = \sigma_0^2 + \cdots + \sigma_m^2$, and $p_k = \sigma_k^2/\sigma^2$. At this stage it is more convenient to assume $\omega_0 = 0$ and write (7.2) in the symmetric form

$$R(v) = \sigma^2 \sum_{k=-m}^{m} q_k \cos v\omega_k,$$

where $q_0 = p_0$, and for $k = 1, \ldots, m$, $q_k = \frac{1}{2}p_k$, $q_{-k} = q_k$, $\omega_{-k} = -\omega_k$. As in Section 1 this suggests the generalization to

$$R(v) = \sigma^2 \int_{-\pi}^{\pi} \cos v\omega \, dF(\omega), \tag{7.3}$$

where F is a cumulative distribution function of a symmetric random variable having possible values in $[-\pi, \pi]$. In the first part of this section we will show that every covariance function may be written in the form of (7.3) for some unique probability distribution function $F(\omega)$. This function thus associated with the covariance function is called the *spectral distribution function* of the process.

In general, it is possible to develop from (7.3) a canonical representation for the process $\{X_n\}$. To sketch the thinking, suppose $\{Z^{(1)}(\omega),\ 0 \le \omega \le \pi\}$ and $\{Z^{(2)}(\omega),\ 0 \le \omega \le \pi\}$ are two stochastic processes, uncorrelated with each other and having uncorrelated increments. Suppose further that $Z^{(i)}(0) = 0$ and that the $Z^{(i)}$ processes have the common variance function

$$\begin{aligned} E[Z^{(i)}(\omega)^2] &= \sigma^2[F(\omega) - F(-\omega)], \\ &= 2\sigma^2[F(\omega) - F(0)], \qquad 0 \le \omega \le \pi. \end{aligned}$$

For expository convenience in this preliminary sketch we have assumed that $F(\omega)$ is continuous.

Let $0 = \omega_0 < \omega_1 < \cdots < \omega_m < \omega_{m+1} = \pi$ be given, and set $\Delta Z_k^{(i)} = Z^{(i)}(\omega_{k+1}) - Z^{(i)}(\omega_k)$. Since nonoverlapping increments are uncorrelated and the variance function is F, we have

$$E[\Delta Z_k^{(i)} \Delta Z_l^{(j)}] = 0, \qquad \text{if} \quad i \ne j, \quad \text{or} \quad k \ne l,$$

and

$$\frac{1}{\sigma^2} E[(\Delta Z_k^{(i)})^2] = 2\{F(\omega_{k+1}) - F(\omega_k)\} = 2\Delta F_k.$$

With $A_k = \Delta Z_k^{(1)}$ and $B_k = \Delta Z_k^{(2)}$, Eq. (7.1) becomes

$$\sum_{k=0}^{m} \cos n\omega_k \Delta Z_k^{(1)} + \sum_{k=0}^{m} \sin n\omega_k \Delta Z_k^{(2)},$$

the Riemann–Stieltjes approximating sum for the integral

$$\int_0^\pi \cos n\omega \, dZ^{(1)}(\omega) + \int_0^\pi \sin n\omega \, dZ^{(2)}(\omega).$$

The approximating sums converge in the mean square sense yielding the required representation. The corresponding development for Gaussian processes is given in the next section.

SPECTRAL REPRESENTATION OF THE COVARIANCE FUNCTION

Let $\{X_n; n = 0, \pm 1, \pm 2, \ldots\}$ be a covariance stationary process having a mean of zero and a known covariance function. By multiplying each observation by a constant, if necessary, we may assume the constant variance of the process is one. Thus we are studying a covariance stationary process $\{X_n\}$ whose covariance function

$$R(v) = E[X_n X_{n+v}], \qquad v = 0, \pm 1, \pm 2, \ldots,$$

satisfies $R(0) = 1$.

Before presenting the first theorem, we give two brief definitions. First, a real-valued function $R(v)$, $v = 0, \pm 1, \pm 2, \ldots$, is called *positive semi-definite* if for all $k = 1, 2, \ldots$ and all real numbers $\alpha_1, \ldots, \alpha_k$,

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j R(i-j) \geq 0. \tag{7.4}$$

Second, a random variable W, or its distribution function F, is said to be *symmetric* if W has the same distribution as $-W$. Thus, F is symmetric if $F(\omega) = 1 - F(-\omega)$ at all points of continuity ω.

Theorem 7.1. *Let $R(v)$, $v = 0, \pm 1, \pm 2, \ldots$, be given. The following statements are equivalent:*

(a) *$R(v)$ is the covariance function of a real-valued covariance stationary process having mean zero and variance one;*

(b) *$R(0) = 1$, $R(v) = R(-v)$, for $v = 0, 1, \ldots$, and $R(v)$ is positive semi-definite;*

(c) *There exists a symmetric probability distribution F on $[-\pi, \pi]$ for which*

$$R(v) = \int_{-\pi}^{\pi} \cos v\omega \, dF(\omega), \qquad v = 0, 1, \ldots. \tag{7.5}$$

Proof. (a) $\Rightarrow$ (b). Let us suppose $R(v)$ is the covariance function of a covariance stationary process $\{X_n\}$ having mean zero and variance one. Clearly, $R(0) = E[X_n^2] = 1$, and $R(v) = E[X_n X_{n+v}] = E[X_{n+v} X_n] = R(-v)$.

To show $R(v)$ is positive semidefinite, let $\alpha_1, \ldots, \alpha_k$ be given, and compute

$$\begin{aligned} 0 &\leq E\left[\left|\sum_{i=1}^{k} \alpha_i X_{n+i}\right|^2\right] \\ &= E\left[\left(\sum_{i=1}^{k} \alpha_i X_{n+i}\right)\left(\sum_{j=1}^{k} \alpha_j X_{n+j}\right)\right] \\ &= E\left[\sum_{i=1}^{k}\sum_{j=1}^{k} \alpha_i \alpha_j X_{n+i} X_{n+j}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j E[X_{n+i} X_{n+j}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j R(i-j). \end{aligned}$$

Thus, (a) implies (b).

(b) $\Rightarrow$ (a). Let us suppose we are given a function $R(v)$ having the properties listed in (b), and set

$$a_{ij} = R(i-j).$$

Then, for any $k = 1, 2, \ldots$, the matrix $A = \|a_{ij}\|_{i,j=1}^{k}$ is symmetric and positive semidefinite. According to our remarks on the multivariate normal distribution in Chapter 1, we may associate with any such matrix A a multivariate normal distribution with zero mean and for which A is the covariance matrix. This procedure describes in a consistent way the distribution of any finite set $(X_{n+1}, \ldots, X_{n+k})$ and thus describes a distribution of the process $\{X_n; n = 0, \pm 1, \ldots\}$. It is easily seen that R is the covariance function of this process, and thus, for every function R with the properties listed in (b), there is at least one covariance stationary process having R as its covariance function.

(c) $\Rightarrow$ (b). We suppose

$$R(v) = \int_{-\pi}^{\pi} \cos v\omega \, dF(\omega) = E[\cos vW],$$

where W is a symmetric random variable having distribution function F. Easily $R(0) = 1$, and, since $\cos v\omega = \cos(-v\omega)$, $R(v) = R(-v)$. We need

only show that $R(v)$ is positive semidefinite. We use the identity $\cos(x-y)=\cos x\cos y+\sin x\sin y$ to write

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i\alpha_j R(i-j)&=\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i\alpha_j E[\cos(i-j)W]\\&=E\left[\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i\alpha_j\cos(iW-jW)\right]\\&=E\left[\left|\sum_{i=1}^{k}\alpha_i\cos iW\right|^2+\left|\sum_{i=1}^{k}\alpha_i\sin iW\right|^2\right]\geq 0.\end{aligned}$$

Thus (c) implies (b).

(b) $\Rightarrow$ (c). Let R be given. For any given n and ω, use property (7.4) with $\alpha_j=\cos j\omega$ to conclude

$$0\leq\sum_{i=1}^{n}\sum_{j=1}^{n}R(i-j)\cos i\omega\cos j\omega.$$

Similarly, with $\alpha_i=\sin i\omega$, we have

$$0\leq\sum_{i=1}^{n}\sum_{j=1}^{n}R(i-j)\sin i\omega\sin j\omega.$$

We add these inequalities and use the trigonometric identity $\cos(x-y)=\cos x\cos y+\sin x\sin y$ to infer

$$\begin{aligned}0&\leq\frac{1}{2\pi n}\sum_{i=1}^{n}\sum_{j=1}^{n}R(i-j)\cos(i-j)\omega\\&=\frac{1}{2\pi n}\sum_{v=-n+1}^{n-1}(n-|v|)\,R(v)\cos v\omega.\end{aligned}$$

Thus, if we define

$$f_n(\omega)=\frac{1}{2\pi n}\sum_{v=-n+1}^{n-1}(n-|v|)\,R(v)\cos v\omega,$$

for $-\pi\leq\omega\leq\pi$, then

$$f_n(\omega)\geq 0.$$

Next we compute,

$$\begin{aligned}\int_{-\pi}^{\pi}f_n(\omega)\,d\omega&=\frac{1}{2\pi n}\sum_{v=-n+1}^{n-1}(n-|v|)\left\{\int_{-\pi}^{\pi}\cos v\omega\,d\omega\right\}R(v)\\&=\frac{1}{2\pi n}(n)(2\pi)R(0)=1.\end{aligned}$$

Thus, for every n, f_n is a probability density function on $[-\pi, \pi]$. Observe from its definition that f_n is symmetric:

$$f_n(\omega) = f_n(-\omega).$$

Let F_n be the cumulative distribution corresponding to the density function f_n. Using the definition for f_n, we compute

$$\begin{aligned}
\int_{-\pi}^{\pi} \cos v\omega \, dF_n(\omega) &= \int_{-\pi}^{\pi} \cos v\omega f_n(\omega) \, d\omega \\
&= \int_{-\pi}^{\pi} \cos v\omega \left\{ \frac{1}{2\pi n} \sum_{k=-n+1}^{n-1} (n - |k|) \cos k\omega R(k) \right\} d\omega \\
&= \sum_{k=-n+1}^{n-1} \left(1 - \frac{|k|}{n}\right) \frac{1}{2\pi} \left\{ \int_{-\pi}^{\pi} \cos v\omega \cos k\omega \, d\omega \right\} R(k) \\
&= \sum_{k=-n+1}^{n-1} \left(1 - \frac{|k|}{n}\right) \delta_{vk} R(k) \\
&= \left(1 - \frac{|v|}{n}\right) R(v),
\end{aligned}$$

where δ_{vk} is one if $v = k$, and zero otherwise. We have thus defined a sequence of distribution functions F_n for which the result (7.5) is "approximately" true, in the sense that

$$\left(1 - \frac{|v|}{n}\right) R(v) = \int_{-\pi}^{\pi} \cos v\omega \, dF_n(\omega).$$

From the Helly–Bray lemma (Chapter 1), there exists a cumulative distribution function F and a subsequence $\{F_{n_k}\}$ such that, for every bounded continuous function h,

$$\lim_{k \to \infty} \int_{-\pi}^{\pi} h(\omega) \, dF_{n_k}(\omega) = \int_{-\pi}^{\pi} h(\omega) \, dF(\omega).$$

We apply this with $h(\omega) = \cos v\omega$ to get

$$\begin{aligned}
\int_{-\pi}^{\pi} \cos v\omega \, dF(\omega) &= \lim_{k \to \infty} \int_{-\pi}^{\pi} \cos v\omega \, dF_{n_k}(\omega) \\
&= \lim_{k \to \infty} \left(1 - \frac{|v|}{n_k}\right) R(v) \\
&= R(v), \qquad \text{for} \quad v = 0, \pm 1, \ldots,
\end{aligned}$$

which completes the proof of the theorem. ■

According to Eq. (7.5), the covariance function $R(v)$ can easily be determined from the spectral distribution function F. In fact,

$$R(v) = \int_{-\pi}^{\pi} \cos v\omega \, dF(\omega)$$

states that $R(v)$ is the Fourier–Stieltjes cosine transformation of the distribution $F(\omega)$. Conversely, in general the spectral distribution function F can be determined from the covariance function $R(v)$ by computing an inverse cosine transformation. We content ourselves with stating without proof the results in an important case in which $F(\omega)$ is differentiable with *spectral density function*

$$f(\omega) = \frac{dF(\omega)}{d\omega}, \qquad \text{for} \quad -\pi < \omega < \pi.$$

In this case,

$$R(v) = \int_{-\pi}^{\pi} \cos v\omega f(\omega) \, d\omega$$

is the cosine transformation of the spectral density function $f(\omega)$.

Theorem 7.2. *Let $F(\omega)$ be the spectral distribution function corresponding to a covariance function $R(v)$. If*

$$\sum_{v=0}^{\infty} |R(v)| < \infty,$$

then $F(\omega)$ is differentiable with derivative

$$f(\omega) = F'(\omega), \qquad -\pi < \omega < \pi.$$

In this case,

$$f(\omega) = \frac{1}{\pi}\left\{\frac{1}{2} R(0) + \sum_{v=1}^{\infty} \cos v\omega R(v)\right\}$$

$$= \frac{1}{2\pi} \sum_{v=-\infty}^{+\infty} R(v) \cos v\omega$$

$$= \frac{1}{2\pi} \sum_{v=-\infty}^{+\infty} e^{iv\omega} R(v).$$

COMPLEX-VALUED PROCESSES

To further the theory and its applications, it is useful to have a theory covering complex-valued stationary processes. As a particular example, in many areas of communication theory it is natural to represent the

instantaneous voltage and current of an alternating current signal by a complex number. Fortunately, it is possible to achieve this desired generality in a very natural way.

Let $X = X_1 + iX_2$, and $Y = Y_1 + iY_2$ be complex random variables. Then we define, as anticipated, the expectation

$$m_X = E[X] = E[X_1] + iE[X_2],$$

but for the covariance, we take

$$\operatorname{cov}[X, Y] = E[(X - m_X)(\overline{Y - m_Y})],$$

where the bar signifies the complex conjugate

$$\bar{Y} = Y_1 - iY_2.$$

Thus the covariance function of a complex covariance stationary process $\{X_n\}$ having zero mean is the complex-valued function

$$R(v) = E[X_n \bar{X}_{n+v}], \qquad v = 0, \pm 1, \pm 2, \ldots.$$

As with real-valued processes, $R(0)$ is real, and

$$R(0) \geq |R(v)|, \qquad v = 0, \pm 1, \pm 2, \ldots.$$

But symmetry now takes the Hermitian form

$$R(v) = \overline{R(-v)},$$

and positive semidefiniteness is the property

$$\sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \bar{\alpha}_k R(j-k) \geq 0,$$

for all $n = 1, 2, \ldots$ and complex numbers $\alpha_1, \ldots, \alpha_n$.

The following theorem is the generalization of Theorem 7.1. The proof parallels that of the earlier theorem in every step and is omitted.

Theorem 7.3. *Let the complex-valued function $R(v)$, $v = 0, \pm 1, \ldots$, be given. The following statements are equivalent:*

(a) *$R(v)$ is the covariance function of a complex-valued covariance stationary stochastic process having zero mean and unit variance;*

(b) *$R(0) = 1$, $R(v) = \overline{R(-v)}$, for $v = 0, 1, \ldots$, and $R(v)$ is positive semidefinite;*

(c) *There exists a probability distribution F (not necessarily symmetric) on $[-\pi, \pi]$ for which*

$$R(v) = \int_{-\pi}^{\pi} e^{iv\omega}\, dF(\omega).$$

8: *Gaussian Systems*

In this section we will construct a real-valued Gaussian stationary process $\{X_n\}$ having symmetric spectral distribution function $F(\omega)$, $-\pi \leq \omega \leq \pi$, through the formula (We use the notation $Z^{(i)}(d\omega) = dZ^{(i)}(\omega)$ interchangeably.)

$$X_n = \int_0^\pi \cos n\omega Z^{(1)}(d\omega) + \int_0^\pi \sin n\omega Z^{(2)}(d\omega), \tag{8.1}$$

where $\{Z^{(i)}(\omega);\ 0 \leq \omega \leq \pi\}$, $i = 1, 2$, are Gaussian processes, independent of one another and having independent increments, and where

$$E[\{Z^{(i)}(\omega_2) - Z^{(i)}(\omega_1)\}^2] = 2\{F(\omega_2) - F(\omega_1)\},$$
$$\text{for} \quad i = 1, 2, \quad \text{and} \quad 0 \leq \omega_1 \leq \omega_2 \leq \pi.$$

In general, a representation of the form (8.1) exists for every covariance stationary process, although $Z^{(i)}$ will not be Gaussian if $\{X_n\}$ is not. The simple result we present is only a sample of what can be done.

GAUSSIAN SYSTEMS

Let T be an abstract set and $\{X(t);\ t \in T\}$ a stochastic process. We call $\{X(t);\ t \in T\}$ a *Gaussian system* or *Gaussian process* if, for every $n = 1, 2, \ldots$ and every finite subset $\{t_1, \ldots, t_n\}$ of T, the random vector $(X(t_1), \ldots, X(t_n))$ has a multivariate normal distribution. Equivalently, we may require for every n that every linear combination

$$\alpha_1 X(t_1) + \cdots + \alpha_n X(t_n), \qquad \alpha_i \quad \text{real},$$

have a univariate normal distribution. Every Gaussian system is described uniquely by its two parameters, the mean and covariance function, given respectively by

$$\mu(t) = E[X(t)], \qquad t \in T,$$

and

$$\Gamma(t_1, t_2) = E[\{X(t_1) - \mu(t_1)\}\{X(t_2) - \mu(t_2)\}], \qquad t_i \in T.$$

The covariance function is positive definite, i.e., for every $n = 1, 2, \ldots$, real numbers $\alpha_1, \ldots, \alpha_n$ and elements $t_1, \ldots, t_n$ in T,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j \Gamma(t_i, t_j) \geq 0. \tag{8.2}$$

We need only compute the variance of $\sum_{i=1}^{n} \alpha_i \{X(t_i) - \mu(t_i)\}$ to verify this.

In fact,

$$0 \le E\left[\left\{\sum_{i=1}^{n} \alpha_i(X(t_i) - \mu(t_i))\right\}^2\right]$$
$$= \sum_{i,\,j=1}^{n} \alpha_i \alpha_j E[\{X(t_i) - \mu(t_i)\}\{X(t_j) - \mu(t_j)\}]$$
$$= \sum_{i,\,j=1}^{n} \alpha_i \alpha_j \Gamma(t_i, t_j).$$

Conversely, given an arbitrary mean $\mu(t)$ and positive definite covariance function $\Gamma(t_1, t_2)$, there exists a corresponding Gaussian system. To convince yourself of this, associate to every finite set $\{X(t_1), \ldots, X(t_n)\}$, $t_i \in T$, a multivariate normal distribution having mean vector $\{\mu(t_1), \ldots, \mu(t_n)\}$ and covariance matrix $\|\Gamma(t_i, t_j)\|_{i,\,j=1}^{n}$. This prescribes a consistent set of finite-dimensional distributions for the process $\{X(t); t \in T\}$, and according to Chapter 1, p. 32, this is enough.

GAUSSIAN RANDOM MEASURE

Let E be a subset of a finite-dimensional space and $g(x)$ a nonnegative function on E for which $\int_E g(x)\,dx < \infty$. For every subset A of E, define

$$m(A) = \int_A g(x)\,dx.$$

We claim that the expression

$$\Gamma(A, B) = m(A \cap B)$$

defines a positive definite function with parameter index consisting of the subsets of E. Almost exactly as before, we compute

$$0 \le \int_E \left\{\sum_{i=1}^{n} \alpha_i I_{A_i}(x)\right\}^2 g(x)\,dx$$
$$= \sum_{i,\,j=1}^{n} \alpha_i \alpha_j \int_E I_{A_i}(x) I_{A_j}(x) g(x)\,dx \tag{8.3}$$
$$= \sum_{i,\,j=1}^{n} \alpha_i \alpha_j m(A_i \cap A_j),$$

where

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases}$$

The Gaussian system $\{Z(A);\ A \subset E\}$ having mean zero and covariance $\Gamma(A, B) = m(A \cap B)$, $A, B \subset E$ is called the *Gaussian random measure* induced by m. Even if $\int_E g(x)\,dx = \infty$, we may define $Z(A)$, not for all subsets A of E, but for all sets A for which $\int_A g(x)\,dx < \infty$.

Suppose E comprises all of a certain finite-dimensional space and $g(x) = 1$ for all x. The corresponding Gaussian random measure $\{Z(A)\}$ is a stationary process whose index set is a family of subsets. The constancy of g ensures that $(Z(A_1 + x), \ldots, Z(A_n + x))$ has the same distribution as $(Z(A_1), \ldots, Z(A_n))$, where $A_1 + x$ is the set A_1 translated by x, i.e., $A_1 + x = \{y : y = x + z \text{ for some } z \in A_1\}$.

Let us look at $E = [a, b]$, where $0 \leq a \leq b < \infty$, and suppose $G(x)$ is a nondecreasing bounded function for $a \leq x \leq b$. Write

$$G(I) = G(x_2) - G(x_1),$$

whenever $I = (x_1, x_2]$, $a \leq x_1 < x_2 \leq b$. We have made G into a nonnegative function of half-open intervals $I = (x_1, x_2] \subset E$. Set $\Gamma(I_1, I_2) = G(I_1 \cap I_2)$. We claim $\Gamma(I_1, I_2)$ is positive semidefinite. When G has a derivative g, the proof is exactly as in (8.3). The general case follows the same recipe, provided we replace $g(x)\,dx$ by the increment $dG(x)$ and use the Riemann–Stieltjes theory of integration mentioned in Chapter 1.

Let $Z(I)$ be a Gaussian random measure on intervals $I = (x, y]$, $a \leq x < y \leq b$, with $E[Z(I)^2] = G(I)$. We will define the integral

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

for continuous functions $f(x)$ by taking a mean square limit of approximating sums

$$\sum_{i=0}^{n-1} f(x_i) Z(I_i),$$

where $I_i = (x_i, x_{i+1}]$, $a = x_0 < x_1 < \cdots < x_n = b$.

Let $\mathscr{P} = \{x_i\}$, $a = x_0 < x_1 < \cdots < x_n = b$, induce a partition of $[a, b]$. We call a partition $\mathscr{P}' = \{x_i'\}$ a *refinement* of $\mathscr{P}$ if every $x_i \in \mathscr{P}$ is also a point in $\mathscr{P}'$. Let

$$\mathscr{I}(f; \mathscr{P}) = \sum_{i=0}^{n-1} f(x_i) Z(I_i), \qquad I_i = (x_i, x_{i+1}].$$

Were $\mathscr{I}(f; \mathscr{P})$ deterministic, we would define an integral by building successive refinements $\mathscr{P}'$ of an arbitrary partition $\mathscr{P}$ and show that the approximating sums $\mathscr{I}(f; \mathscr{P}')$ converge. We follow the same pattern here, but since $\mathscr{I}(f; \mathscr{P})$ is random, we substitute mean square convergence for

the ordinary convergence of real numbers. To be precise, we will show that for every preassigned $\varepsilon > 0$ there exists a partition $\mathscr{P}$, such that for all refinements $\mathscr{P}'$

$$\|\mathscr{I}(f; \mathscr{P}) - \mathscr{I}(f; \mathscr{P}')\| < \varepsilon,$$

where $\|\cdot\|$ is the mean square distance of Section 2. Using the *completeness* property of mean square distance discussed in that section, this will imply the existence of a limit in the mean square sense as the partitions are refined. This limit is defined to be the integral

$$\mathscr{I}(f) = \int_a^b f(x) Z(dx).$$

Recall that if f is continuous it is uniformly continuous on $[a, b]$. We thus may choose a partition $\mathscr{P} = \{x_i\}$ for which the range of f on any subinterval $I_i = (x_i, x_{i+1}]$ is smaller than

$$\delta = \varepsilon / \sqrt{\{G(b) - G(a)\}},$$

for any preassigned $\varepsilon > 0$. Let $\mathscr{P}' = \{x_j'\}$ be a refinement of $\mathscr{P}$. Each interval $I_j' = (x_j', x_{j+1}']$ is contained in some $I_i = (x_i, x_{i+1}]$. Certainly $\varepsilon_j = f(x_i) - f(x_j') < \delta$. Then

$$\begin{aligned}
\|\mathscr{I}(f; \mathscr{P}) - \mathscr{I}(f; \mathscr{P}')\|^2 &= E[\{\mathscr{I}(f; \mathscr{P}) - \mathscr{I}(f; \mathscr{P}')\}^2] \\
&= E\left[\left\{\sum_{i=0}^{n-1} f(x_i) Z(I_i) - \sum_{j=0}^{m-1} f(x_j') Z(I_j')\right\}^2\right] \\
&= E\left[\left\{\sum_{j=0}^{m-1} \varepsilon_j Z(I_j')\right\}^2\right] \\
&= \sum_{i,\, j=0}^{m-1} \varepsilon_i \varepsilon_j G(I_i' \cap I_j') \\
&\leq \delta^2 \sum_{i=0}^{m-1} G(I_j') = \delta^2 \{G(b) - G(a)\} = \varepsilon^2.
\end{aligned}$$

Since ε is arbitrary, the approximating sums converge in the mean square sense as the partitions are refined. The limit achieved is the definition of the integral

$$\mathscr{I}(f) = \int_a^b f(x) Z(dx).$$

It is a random variable defined as a mean square limit.

The integral $\mathscr{I}(\cdot)$ possesses most of the properties of the usual integral. We mention here linearity:

$$\mathscr{I}(\alpha f_1 + \beta f_2) = \alpha \mathscr{I}(f_1) + \beta \mathscr{I}(f_2),$$

for real α, β and continuous f_i, provided we interpret "$=$" as equal in mean square distance,

$$\|\mathscr{I}(\alpha f_1 + \beta f_2) - \alpha \mathscr{I}(f_1) - \beta \mathscr{I}(f_2)\| = 0.$$

Since $\mathscr{I}(f)$ is random, we may compute its mean and variance. Recall from Section 2 that the expectation and covariance are continuous under mean square convergence. Thus

$$E[\mathscr{I}(f)] = \lim E\left[\sum_i f(x_i) Z(I_i)\right] = 0,$$

and

$$\begin{aligned} E[\mathscr{I}(f)\mathscr{I}(h)] &= \lim E\left[\sum_i f(x_i) Z(I_i) \sum_j h(x_j) Z(I_j)\right] \\ &= \lim \sum_{i,j} f(x_i) h(x_j) G(I_i \cap I_j) \\ &= \lim \sum_{i,j} f(x_i) h(x_j) \int I_{I_i}(x) I_{I_j}(x)\, dG(x) \\ &= \lim \int \left\{\sum_i f(x_i) I_{I_i}(x)\right\}\left\{\sum_j h(x_j) I_{I_j}(x)\right\} dG(x) \\ &= \int_a^b f(x) h(x)\, dG(x), \end{aligned}$$

for any continuous $f(x)$ and $h(x)$.

Deceptively simple,

$$E[\mathscr{I}(f)\mathscr{I}(h)] = \int_a^b f(x) h(x)\, dG(x) \tag{8.4}$$

is the most important property of integration with respect to Gaussian random measures.

SPECTRAL REPRESENTATION OF A GAUSSIAN STATIONARY PROCESS.

Let $Z^{(1)}$ and $Z^{(2)}$ be independent Gaussian random measures on $[0, \pi]$ satisfying

$$E[\{Z^{(i)}(I)\}^2] = 2F(I),$$

where $I = (\omega_1, \omega_2]$, $0 \leq \omega_1 < \omega_2 \leq \pi$, and $F(\omega)$, $-\pi \leq \omega \leq \pi$, is a given symmetric distribution function. We assume $F(0) = F(0-)$. The reader may wish to verify that such random measures may be explicitly represented by the formula

$$Z^{(i)}(I) = \sqrt{2}\{B(F(\omega_2)) - B(F(\omega_1))\},$$

where $\{B(t); t \geq 0\}$ is a standard Brownian motion (see Chapter 7). Using the stochastic integral $\mathscr{I}(\cdot)$, define

$$\begin{aligned} X_n &= \int_0^\pi \cos n\omega Z^{(1)}(d\omega) + \int_0^\pi \sin n\omega Z^{(2)}(d\omega), \qquad n = 0, \pm 1, \ldots, \\ &= \mathscr{I}_1(\cos n\omega) + \mathscr{I}_2(\sin n\omega). \end{aligned} \tag{8.5}$$

Then, $E[X_n] = 0$, and, following (8.4),

$$\begin{aligned} &E[X_n X_{n+v}] \\ &\quad = E[\{\mathscr{I}_1(\cos n\omega) + \mathscr{I}_2(\sin n\omega)\}\{\mathscr{I}_1(\cos(n+v)\omega) + \mathscr{I}_2(\sin(n+v)\omega)\}] \\ &\quad = E[\mathscr{I}_1(\cos n\omega)\mathscr{I}_1(\cos(n+v)\omega)] + E[\mathscr{I}_2(\sin n\omega)\mathscr{I}_2(\sin(n+v)\omega)] \\ &\quad = 2\int_0^\pi \cos n\omega \cos(n+v)\omega \, dF(\omega) + 2\int_0^\pi \sin n\omega \sin(n+v)\omega \, dF(\omega) \\ &\quad = \int_{-\pi}^{\pi} \cos v\omega \, dF(\omega). \end{aligned}$$

It follows from Theorem 7.1 that $\{X_n\}$ is a covariance stationary process having the symmetric spectral distribution function F.

We remark that a parallel development exists for complex-valued Gaussian stationary processes in terms of a complex-valued Gaussian random measure.

Equation (8.5) represents the stationary Gaussian process $\{X_n\}$ in terms of the Gaussian random measure $Z^{(i)}$. We state without proof the converse. Let $\{X_n\}$ be a stationary Gaussian process having spectral distribution function F. The formulas

$$Z^{(1)}(\lambda) = \frac{1}{2\pi}\left\{\lambda X_0 + \sum_{n=1}^{\infty} \frac{1}{n}(X_n - X_{-n}) \sin \lambda n\right\},$$

and

$$Z^{(2)}(\lambda) = \frac{1}{2\pi}\sum_{n=1}^{\infty} \frac{1}{n}(X_n + X_{-n}) \cos \lambda n,$$

define independent Gaussian processes having independent increments.

The infinite summations are understood in the mean square sense, which suffices to define all finite-dimensional distributions of $Z^{(i)}(\lambda)$. If $I=(\omega_1, \omega_2]$ and $Z^{(i)}(I) = Z^{(i)}(\omega_2) - Z^{(i)}(\omega_1)$, then

$$E[\{Z^{(i)}(I)\}^2] = F(\omega_2) - F(\omega_1).$$

9: *Stationary Point Processes*

Let $\mathscr{A}$ be a family of subsets of the nonnegative half-line $[0, \infty)$, and suppose that every interval of the form $(t, s]$, $0 \leq t < s$, is a member of $\mathscr{A}$. A point process $\{N(A);\ A \in \mathscr{A}\}$ is said to be *stationary* if, for every real number h, every positive integer k, and every set of intervals

$$(t_1, s_1], \ldots, (t_k, s_k],$$

the k-dimensional random vector

$$(N(t_1, s_1], \ldots, N(t_k, s_k])$$

has the same joint distribution as the vector

$$(N(t_1 + h, s_1 + h], \ldots, N(t_k + h, s_k + h]).$$

Of course, an analogous definition can be formed when $\mathscr{A}$ is a family of subsets of the whole real line, or even a finite-dimensional Euclidean space.

If $\{N(A),\ A \in \mathscr{A}\}$ is a stationary point process, then for each fixed $t, s \geq 0$, the integer-valued stochastic process

$$W(h) = N(t + h, s + h], \qquad h \geq 0,$$

is a stationary process and thus amenable to the techniques of the first eight sections of this chapter. However, the special character of point processes merits a separate study.

Suppose $\{X(t);\ t \geq 0\}$ is a stationary process for which every trajectory $X(t)$ is a continuous function of t. Fix a level u and let $N_u(s, t]$ be the number of times the trajectory $X(t)$ crosses u in the time interval $(s, t]$. Then $\{N_u(s, t];\ 0 \leq s < t < \infty\}$ is a stationary point process. In Section 10 we will compute the mean $E[N_u(0, t]]$ for certain Gaussian stationary processes $\{X(t)\}$. In the remainder of this section, we content ourselves with showing that a stationary renewal process (Section 7 of Chapter 5) induces a stationary point process.

Let F be an arbitrary cumulative distribution function of a nonnegative random variable X having a finite mean

$$\mu = E[X] = \int_0^\infty x\, dF(x)$$
$$= \int_0^\infty [1 - F(y)]\, dy,$$

the last formula resulting by integration by parts (see also Problem 9, Chapter 1). Thus, since it is nonnegative and integrates to one, the function

$$g(y) = \mu^{-1}[1 - F(y)], \qquad y \geq 0,$$

is a probability density function.

Now let $X_0, X_1, X_2, \ldots$ be independent random variables, where X_0 is continuous and has the probability density function g, while for $i \geq 1$, X_i has the distribution function F. Construct a point process by placing a "point" at each of the times $S_n = X_0 + \cdots + X_n$, for $n \geq 0$. More precisely, for any interval $A = (t, s]$, we let $N(A)$ be the number of indices n for which $t < S_n \leq s$.

We claim that the point process $\{N(A), A \in \mathscr{A}\}$, constructed in this manner is a stationary point process. Now it is quite hard to exhibit explicitly the joint distribution of an arbitrary vector $\mathbf{N}(0) = (N(t_1, s_1], \ldots, N(t_k, s_k])$. However, we need not determine this distribution but merely show that it is the same as that for the shifted vector $\mathbf{N}(h) = (N(t_1 + h, s_1 + h], \ldots, N(t_k + h, s_k + h])$. Let us write X for the entire infinite-dimensional vector $(X_0, X_1, \ldots)$, and observe that $\mathbf{N}(0)$ is a vector-valued function of X. Let f denote this function, and let f_h be the vector-valued function that carries X into $\mathbf{N}(h)$. Then we want to show that $\mathbf{N}(0) = f(X)$ and $\mathbf{N}(h) = f_h(X)$ have the same distribution. Formulated this way, the problem is quite difficult.

The trick is to express $\mathbf{N}(h)$ involving the same function f used to produce $\mathbf{N}(0)$, but evaluated at a shifted sequence $X' = (X'_0, X'_1, \ldots)$. We claim

$$\mathbf{N}(h) = f(X'), \tag{9.1}$$

where

$$X'_0 = S_M - h, \qquad X'_k = X_{M+k}, \qquad k \geq 1, \tag{9.2}$$

and where M is determined by

$$S_{M-1} < h \leq S_M. \tag{9.3}$$

Then to show $\mathbf{N}(h) = f(X')$ has the same distribution as $\mathbf{N}(0) = f(X)$, one need only show that the infinite-dimensional vector $X' = (X'_0, X'_1, \ldots)$ has the same distribution as the infinite-dimensional vector $X = (X_0, X_1, \ldots)$, since the function f is the same for both $\mathbf{N}(0)$ and $\mathbf{N}(h)$.

To verify Eq. (9.1) is not hard. We consider only the one-dimensional case in which

$$\mathbf{N}(0) = N(t, s],$$

for some fixed $t < s$. Let $\#\{\ \}$ denote "number of," so that we write

$$\begin{aligned} N(t, s] &= \#\{n : t < S_n \leq s\} \\ &= f(X). \end{aligned}$$

Then

$$\begin{aligned} N(t+h, s+h] &= \#\{n \geq 0 : t+h < S_n \leq s+h\} \\ &= \#\{n \geq M : t+h < S_n \leq s+h\} \\ &= \#\{n \geq M : t < S_n - h \leq s\} \\ &= \#\{m \geq 0 : t < S_{m+M} - h \leq s\} \\ &= \#\{m \geq 0 : t < S'_m \leq s\} \\ &= f(X'), \end{aligned}$$

where, as before, X' and M are given by Eqs. (9.2) and (9.3), and $S'_m = X'_0 + \cdots + X'_m$.

Thus, to verify that $\{N(A);\ A \in A\}$ is a stationary point process, we need only show

$$X' = (S_M - h, X_{M+1}, X_{M+2}, \ldots)$$

has the same distribution as

$$X = (X_0, X_1, X_2, \ldots).$$

Now M is determined by the random variables $X_0, \ldots, X_M$, and thus is independent of $X_{M+1}, X_{M+2}, \ldots$. It follows that the distribution of X_{M+k} is the same as the conditional distribution of X_{M+k} given $M = n$, which, using independence, is the same as that of X_{n+k}, which is the same as that of X_k. Similarly, $X_{M+1}, X_{M+2}, \ldots$ are independent, and independent of $S_M - h$. Thus $X'_1, X'_2, \ldots$ are independent and identically distributed, with cumulative distribution function F, and independent of $X'_0 = S_M - h$. It remains only to show that $X'_0 = S_M - h$ has the probability density function $g(y) = \mu^{-1}[1 - F(y)]$, $y \geq 0$. But this was shown in Section 7 of Chapter 5, where $X'_0 = S_M - h$ was identified as the residual life at time

h. Thus, every stationary renewal process induces an associated stationary point process on the positive real line.

The following is the generalization to the whole real line.

Theorem 9.1. *Let* $\{X_n; n = 0, \pm 1, \pm 2, \ldots\}$ *be independent positive random variables. For* $k = \pm 1, \pm 2, \ldots$, *we suppose the* X_k *have the common probability density function* $f(x)$ *for* $x \geq 0$. *We suppose* X_0 *has the distribution of a sum* $X_0^+ + X_0^-$, *where the joint distribution of* X_0^+, X_0^- *is given by the density function*

$$\mu^{-1} f(x^+ + x^-), \qquad \text{for} \quad x^+ \geq 0, \quad x^- \geq 0.$$

If "points" are placed on the real line at $S_n = X_0^+ + \cdots + X_n$ *and at* $S_{-n} = -X_0^- - \cdots - X_{-n}$, *for* $n = 0, 1, 2, \ldots$, *the resulting point process is stationary.*

10: The Level-Crossing Problem

In this section we suppose $\{X(t); -\infty < t < \infty\}$ is a zero-mean Gaussian stationary process for which every trajectory $X(t)$ is a continuous function of t. We fix a level a and consider the number of times the trajectory $X(t)$ crosses a in the time interval $(0, T]$. This quantity has considerable importance in communication theory as well as arising in a variety of other fields. It is quite difficult to compute the distribution of this random variable, and, indeed, in many cases an explicit form is not known. We confine ourselves to computing the first moment, or mean, and this under the additional condition that

$$\lambda_2 = -\frac{d^2}{dt^2} R(t)\Big|_{t=0} < \infty, \tag{10.1}$$

where $R(t)$ is the covariance function of the process.

The derivation uses several techniques of rather general applicability, which we display as lemmas.

Use the notation

$$N(I) = N(s, t], \qquad I = (s, t], \qquad 0 \leq s < t,$$

for a point process. Say that $\{N(I)\}$ is without multiple events if

$$\lim_{n \to \infty} N(t - (1/n), t] = 0 \quad \text{or} \quad 1, \qquad \text{for all} \quad t. \tag{10.2}$$

The renewal point process of Section 9 furnishes an example.

Lemma 10.1. *Let $\{N(I)\}$ be a point process without multiple events. Fix $T > 0$ and divide the interval $(0, T]$ into n subintervals*

$$I_{ni} = ((i-1)T/n, iT/n], \qquad i = 1, \ldots, n; \quad n = 1, 2, \ldots.$$

Then

$$E[N(0, T]] = \lim_{n\to\infty} \sum_{i=1}^{n} \Pr\{N(I_{ni}) \geq 1\}. \tag{10.3}$$

Proof. Write

$$\chi_{ni} = \begin{cases} 1, & \text{if } N(I_{ni}) \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

and $N_n = \sum_{i=1}^{n} \chi_{ni}$. Then $N_n \leq N_{n+1} \leq N(0, T]$ and, in view of (10.2), $\lim_{n\to\infty} N_n = N(0, T]$. The interchange of limit and expectation may be justified to conclude

$$\lim_{n\to\infty} \sum_{i=1}^{n} \Pr\{N(I_{ni}) \geq 1\} = \lim_{n\to\infty} E[N_n] = E[N(0, T]]. \quad \blacksquare$$

Equation (10.3) expresses the mean number of events in $[0, T]$ in terms of the distributions of events in small intervals. We are interested in events defined by the crossings of a level a by a continuous process $X(t)$. Our next step is to relate the number of crossings in an interval to the process $X(t)$, and to do this we need a crisp definition of "crossing." Fix a. Then $X(t)$ is said to have a *crossing* of a at t_0 if for every positive ε there are points t_1 and t_2 satisfying, on the one hand,

$$|t_i - t_0| < \varepsilon, \qquad i = 1, 2,$$

and, on the other,

$$[X(t_1) - a] \cdot [X(t_2) - a] < 0.$$

Observe, for example, that tangencies are not counted as crossings. Let $N_a(0, T]$ be the number of crossings of a by $X(t)$ during $(0, T]$.

Lemma 10.2. *Let $\{X(t)\}$ be a stochastic process for which $X(t)$ is a continuous function of t and for which $\Pr\{X(t) = a\} = 0$ for every fixed t. Then*

$$E[N_a(0, T]] = \lim_{n\to\infty} \left\{ \sum_{i=1}^{n} \Pr\left\{ X\left(\frac{(i-1)T}{n}\right) < a < X\left(\frac{iT}{n}\right) \right\} + \sum_{i=1}^{n} \Pr\left\{ X\left(\frac{(i-1)T}{n}\right) > a > X\left(\frac{iT}{n}\right) \right\} \right.$$

Proof. This result is nearly the same as Lemma 10.1. Let

$$\chi'_{ni} = \begin{cases} 1, & \text{if } X\left(\frac{(i-1)T}{n}\right) < a < X\left(\frac{iT}{n}\right) \text{ or } X\left(\frac{(i-1)T}{n}\right) \\ & \qquad > a > X\left(\frac{iT}{n}\right), \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, $N'_n = \sum_{i=1}^{n} \chi'_{ni} \leq N_n$, so that

$$\limsup_{n\to\infty} E[N'_n] \leq E[N_a(0, T]],$$

by Lemma 10.1. On the other hand, it is apparent that for each fixed n, $N_n \leq N'_m$ when m is sufficiently large, since in that case, for example, $(X(t_1) - a) < 0$, $(X(t_2) - a) > 0$, for some

$$\frac{(i-1)T}{n} < t_1 < t_2 < \frac{iT}{n},$$

will imply

$$X\left(\frac{(j-1)T}{m}\right) < a < X\left(\frac{jT}{m}\right),$$

for some subinterval I_{mj} of I_{ni}. [We have excluded the zero probability event that $X(t)$ crosses a at a point of the form $t = rT$, where r is a rational number in (0, 1].] Thus

$$\liminf_{m\to\infty} E[N'_m] \geq E[N_n],$$

and

$$\liminf_{m\to\infty} E[N'_m] \geq \liminf_{n\to\infty} E[N_n] = E[N_a(0, T)].$$

This completes the proof. ■

Thus the calculation of the mean number of crossings in an interval may be carried out by studying the simpler events

$$\left\{X\left(\frac{(i-1)T}{n}\right) < a < X\left(\frac{iT}{n}\right)\right\},$$

and

$$\left\{X\left(\frac{(i-1)T}{n}\right) > a > X\left(\frac{iT}{n}\right)\right\}.$$

If the process is stationary, these events have the same probabilities, respectively, as do

$$A(n) = \{X(0) < a < X(T/n)\},$$

and

$$B(n) = \{X(0) > a > X(T/n)\}.$$

Theorem 10.1. *Let $\{X(t);\ t \geq 0\}$ be a zero-mean Gaussian stationary process having every trajectory $X(t)$ continuous in t. Suppose the covariance function $R(t)$ satisfies* (10.1), *and set $\sigma^2 = R(0)$, the variance of $X(t)$. Then the mean number of crossings of level a during $(0, T]$ is given by*

$$E[N_a(0, T]] = \frac{T}{\pi}(\lambda_2/\sigma^2)^{1/2}\exp(-a^2/2\sigma^2).$$

Proof. The final result must be proportional to T, by the stationary property, and thus we need only treat $T = 1$. According to the preliminaries, our task is to compute

$$\begin{aligned} E[N_a(0, 1]] &= \lim_{n\to\infty} n[\Pr\{A(n)\} + \Pr\{B(n)\}] \\ &= \lim_{n\to\infty} 2^n[\Pr\{A(2^n)\} + \Pr\{B(2^n)\}]. \end{aligned}$$

Now $A(2^n)$ is determined by $X(0)$ and $X(2^{-n})$, whose joint distribution is normal with mean zero, variance $\sigma^2 = R(0)$, and correlation coefficient $\rho(2^n) = R(2^{-n})/R(0)$. (See Chapter 1 for a review of the bivariate normal distribution.) Let

$$\zeta_n = 2^n[X(2^{-n}) - X(0)].$$

The pair $X(0)$, ζ_n has a bivariate normal distribution, and a straightforward computation shows the mean is zero and the covariance matrix is:

$$\begin{Vmatrix} R(0) & -2^n[R(0) - R(2^{-n})] \\ -2^n[R(0) - R(2^{-n})] & 2^n\{2^n[R(0) - R(2^{-n})] - 2^n[R(-2^{-n}) - R(0)]\}. \end{Vmatrix}$$

Observe the use of the symmetry $R(-2^{-n}) = R(2^{-n})$ in deriving the entries of the matrix. Using the assumption (10.1), this matrix converges, as $n \to \infty$, to

$$\begin{Vmatrix} R(0) & R'(0) \\ R'(0) & R''(0) \end{Vmatrix} = \begin{Vmatrix} \sigma^2 & 0 \\ 0 & \lambda_2 \end{Vmatrix} \tag{10.4}$$

where $R'(t)$ and $R''(t)$ are the first and second derivatives, respectively, of $R(t)$. The symmetry of $R(t)$ and the assumed existence of $R''(0) < \infty$

imply $R'(0) = 0$. Let $p_n(x, z)$ denote this bivariate normal density of $X(0)$ and ζ_n. Note that $B(2^n)$ may be described as

$$\begin{aligned} B(2^n) &= \{X(0) > a > X(2^{-n})\} \\ &= \{a < X(0) < a - 2^{-n}\zeta_n\}. \end{aligned}$$

Now compute

$$\begin{aligned} 2^n \Pr\{B(2^n)\} &= 2^n \Pr\{a < X(0) < a - 2^{-n}\zeta_n\} \\ &= 2^n \int_{-\infty}^{0} \int_{a}^{a-2^{-n}z} p_n(x, z)\, dx\, dz \\ &= \int_{-\infty}^{0} \int_{0}^{-z} p_n(a + 2^{-n}x, z)\, dx\, dz. \end{aligned}$$

Using the explicit form of the bivariate normal distribution given in Chapter 1 and the convergence of the covariance of $X(0)$ and ζ_n to that given in (10.4), we deduce

$$\begin{aligned} \lim_{n\to\infty} p_n(a + 2^{-n}x, z) &= \frac{1}{2\pi\sigma\sqrt{\lambda_2}} \exp\left\{-\frac{1}{2}\left[\left(\frac{a}{\sigma}\right)^2 + \frac{z^2}{\lambda_2}\right]\right\} \\ &= \frac{1}{\sigma}\phi\left(\frac{a}{\sigma}\right) \cdot \frac{1}{\lambda_2^{1/2}}\phi\left(\frac{z}{\lambda_2^{1/2}}\right), \end{aligned}$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

is the standard normal density.

Thus

$$\begin{aligned} \lim_{n\to\infty} 2^n \Pr\{B(2^n)\} &= \int_{-\infty}^{0} \int_{0}^{-z} \frac{1}{\sigma}\phi\left(\frac{a}{\sigma}\right) \frac{1}{\sqrt{\lambda_2}}\phi\left(\frac{z}{\sqrt{\lambda_2}}\right) dx\, dz \\ &= \frac{1}{\sigma}\phi\left(\frac{a}{\sigma}\right) \int_{0}^{\infty} \frac{z}{\sqrt{\lambda_2}}\phi\left(\frac{z}{\sqrt{\lambda_2}}\right) dz \\ &= \frac{\sqrt{\lambda_2}}{\sigma}\phi\left(\frac{a}{\sigma}\right) \int_{0}^{\infty} y\phi(y)\, dy \\ &= \frac{\sqrt{\lambda_2}}{\sigma}\phi\left(\frac{a}{\sigma}\right)\frac{1}{\sqrt{2\pi}} \\ &= \frac{\sqrt{\lambda_2}}{\sigma 2\pi} \exp\left\{-\frac{1}{2}\left(\frac{a}{\sigma}\right)^2\right\}. \end{aligned}$$

We obtain the same result when we compute $\lim_{n\to\infty} 2^n \Pr\{A(2^n)\}$.

Hence

$$E[N_a(0, 1)] = 2 \lim_{n\to\infty} 2^n \Pr\{B(2^n)\}$$
$$= \frac{\sqrt{\lambda_2}}{\pi\sigma} \exp\left\{-\frac{1}{2}\left(\frac{a}{\sigma}\right)^2\right\},$$

as claimed. ■

As a refinement, say that $X(t)$ has an upcrossing of the level a at t_0 if, for some $\varepsilon > 0$, $X(t) < a$ for $t_0 - \varepsilon < t < t_0$, and $X(t) > a$ for $t_0 < t < t_0 + \varepsilon$. Let $U_a(0, T]$ be the number of upcrossings during $(0, T]$. Then

$$E[U_a(0, T]] = \lim_{n\to\infty} 2^n \Pr\{A(2^n)\} = \frac{\sqrt{\lambda_2}}{2\pi\sigma} \exp(-a^2/2\sigma^2).$$

To close this section and this chapter, we state one of the more recent results in this fascinating area of probability. Our aim is to whet the student's appetite for further reading. As the level a increases, the upcrossings become rarer, thus raising the possibility of Poisson-like behavior. Of course, as a increases, $U_a(0, T]$ will tend to become smaller, and thus we require normalization. Let

$$f(a) = \frac{\sqrt{\lambda_2}}{2\pi\sigma} \exp(-a^2/2\sigma^2),$$

and

$$N_a(t) = U_a(0, t/f(a)).$$

Observe that $E[N_a(t)] = t$. Here is the result.

Theorem 10.2. *With the above notation, let $X(t)$ be a stationary Gaussian process, with continuous trajectories, and covariance function $R(t)$, where $\lambda_2 = R''(0) < \infty$. Suppose that either:* (i) $R(t) \log t \to 0$ *as* $t \to \infty$, *or* (ii) $\int_0^\infty R(s)^2\, ds < \infty$. *Then the distribution of* $\{N_a(t);\ t \geq 0\}$ *converges to that of a Poisson process as* $a \to \infty$.

In thinking about why this might be so, observe that (i) and (ii) imply an asymptotic independence that will reflect itself in the independent increments of the Poisson process.

Elementary Problems

Exercises 1–5 all fall into the following context. Let $\{X_n\}$ and $\{Y_n\}$ be jointly distributed zero-mean covariance stationary processes having covariance functions $R_X(v)$ and $R_Y(v)$, respectively. Let $R_{XY}(v) = E[X_n Y_{n+v}]$, $v = 0$,

± 1, ..., be the cross covariance function. Finally, let $\{\xi_n\}$ be a zero-mean covariance stationary process having covariance function $R_\xi(v)$ and being uncorrelated with $\{X_n\}$ or $\{Y_n\}$. "Best" predictor, estimator, etc., is in the sense of minimum mean square error.

1. (a) Find the best predictor of X_{n+1} of the form $\hat{X}^{(1)}_{n+1} = aX_n$, where a is a constant to be chosen.

Solution: $a^* = R_X(1)/R_X(0)$.

(b) Find the best predictor of X_{n+1} of the form $\hat{X}^{(2)}_{n+1} = aX_n + bX_{n-1}$, where a, b are constants.

Solution:

$$a^* = \frac{1}{\Delta}[R_X(1)R_X(0) - R_X(1)R_X(2)],$$

$$b^* = \frac{1}{\Delta}[R_X(0)R_X(2) - R_X(1)^2],$$

where $\Delta = R_X(0)^2 - R_X(1)^2$.

(c) Express the improvement in mean square predictor error of $\hat{X}^{(2)}_{n+1}$ over $\hat{X}^{(1)}_{n+1}$ in terms of $R_X(v)$.

Solution:

$$\text{Difference in MSE} = \frac{1}{R_X(0)}\left\{R_X(2) - \frac{R_X(1)^2}{R_X(0)}\right\}^2.$$

2. (a) Find the best estimator of X_n of the form $\hat{X}^{(1)}_n = aY_n$, where a is a constant to be chosen.

Solution: $a^* = R_{XY}(0)/R_Y(0)$.

(b) Find the best estimator of X_n of the form $\hat{X}^{(2)}_n = aY_n + bY_{n-1}$, where a and b are constants.

Solution:

$$a^* = \frac{1}{\Delta}[R_{XY}(0)R_Y(0) - R_{XY}(1)R_Y(1)],$$

$$b^* = \frac{1}{\Delta}[R_Y(0)R_{XY}(1) - R_Y(1)R_{XY}(0)],$$

where $\Delta = R_Y(0)^2 - R_Y(1)^2$.

3. Interpret X_n as a signal and ξ_n as a noise. We observe $Z_n = X_n + \xi_n$. Find the best estimator of X_n of the form $\hat{X}_n = aZ_n + bZ_{n-1}$, where a and b are constants to be chosen.

Solution:

$$a^* = \frac{1}{\Delta}[R_X(0)\{R_X(0) + R_\xi(0)\} - R_X(1)\{R_X(1) + R_\xi(1)\}],$$

$$b^* = \frac{1}{\Delta}[\{R_X(0) + R_\xi(0)\}R_X(1) - \{R_X(1) + R_\xi(1)\}R_X(0)],$$

where $\Delta = \{R_X(0) + R_\xi(0)\}^2 - \{R_X(1) + R_\xi(1)\}^2$.

4. Find the best interpolator for X_{n+k} of the form

$$\hat{X}_{n+k} = aX_n + bX_{n+N},$$

where $1 \leq k \leq N$ are fixed and a, b are constants to be chosen.

Solution:

$$a^* = \frac{1}{\Delta}[R_X(k)R_X(0) - R_X(N)R_X(N-k)],$$

$$b^* = \frac{1}{\Delta}[R_X(0)R_X(N-k) - R_X(k)R_X(N)],$$

where $\Delta = R_X(0)^2 - R_X(N)^2$.

5. Fix $N \geq 1$ and set $Z_n = \sum_{k=0}^{N} X_{n+k}$. Find the best estimator of Z_n of the form

$$\hat{Z}_n = aX_n + bX_{n+N},$$

where a, b are constants to be chosen.

6. For $n = 1, 2, \ldots$ let $X_n = \cos(nU)$ where U is uniformly distributed over the interval $[-\pi, \pi]$. Verify that $\{X_n\}$ is covariance stationary but not strictly stationary.

Hint: For the first part use the trigonometric identity

$$\cos x \cos y = \tfrac{1}{2}[\cos(x+y) + \cos(x-y)].$$

Evaluate

$$E[\cos vU] = \begin{cases} 1 & \text{if} \quad v = 0 \\ 0 & \text{if} \quad v = 1, 2, \ldots \end{cases}$$

by symmetry. For the second part, use the same approach to determine that the third product moment $E[X_n X_{n+v} X_{n+v+h}]$ depends on n.

7. Suppose $\{B(t);\ t \geq 0\}$ is a standard Brownian motion process. Compute the covariance function for $X(t) = B(t+1) - B(t)$, $t \geq 0$, and establish that $X(t)$ is strictly stationary.

Answer:

$$R(v) = \begin{cases} 1 - |v|, & \text{for} \quad |v| \leq 1, \\ 0, & \text{for} \quad |v| > 1. \end{cases}$$

8. Compute the covariance function for $X_n = \sqrt{2}\, A \sin(\omega n + U)$ where A is a random variable with mean zero and variance σ^2, and U, independent of A, is uniformly distributed over the interval $[0, 2\pi)$. Assume ω is a constant satisfying $0 \leq \omega < 2\pi$.

9. Suppose $\{X_n;\ n = 0, 1, \ldots\}$ is a zero mean stationary process which is both Gaussian and Markov. Demonstrate that the covariance function must be of the form $R(v) = \sigma^2 \lambda^{|v|}$ for some fixed λ satisfying $|\lambda| \leq 1$.

10. Find the spectral density function corresponding to the covariance function $R(0) = 1$ and $R(v) = \alpha \gamma^{|v|}$, $v = \pm 1, \pm 2, \ldots$, where $0 < \alpha < 1$ and $|\gamma| < 1$.

Hint: Write $R(v) = R_1(v) + R_2(v)$ where

$$R_1(v) = \begin{cases} 1 - \alpha & \text{for} \quad v = 0 \\ 0 & \text{for} \quad v \neq 0. \end{cases}$$

and $R_2(v) = \alpha \gamma^{|v|}$ for all v.

11. Suppose $\hat{X}_n = a_1 X_{n-1} + a_2 X_{n-2}$ is an optimal linear predictor for X_n given the entire past of a covariance stationary process $\{X_n\}$. What is the spectral density function?

Hint:

$$\xi_n = X_n - \hat{X}_n = X_n - a_1 X_{n-1} - a_2 X_{n-2}$$

implies

$$\sigma_\xi^2 f_\xi(\lambda) = \sigma_X^2 |1 - a_1 e^{i\lambda} - a_2 e^{2i\lambda}|^2 f_X(\lambda)$$

But we know, since $\hat{X}_n$ is an optimal predictor, $\{\xi_n\}$ are uncorrelated and $f_\xi(\lambda) = \dfrac{1}{2\pi}$, $-\pi \leq \lambda \leq \pi$. Solve for f_X.

Problems

1. Let $\{\xi_n\}$ be independent and identically distributed random variables having mean zero and variance σ^2. Let $\{a_n\}$ be a real sequence. Prove that $X = \sum_{n=0}^{\infty} a_n \xi_n$ converges in mean square whenever $\sum_{n=0}^{\infty} a_n^2 < \infty$ [In particular, $\sum (1/n)\xi_n$ converges!]. Let $\{\eta_n\}$ be a zero-mean covariance stationary process

having covariance function $R(v)$. Show that $\sum_{n=0}^{\infty} a_n \eta_n$ converges in mean square whenever

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} |a_i a_j R(i-j)| < \infty.$$

2. Let $X, X_1, X_2, \ldots$ be random variables having finite second moments. Show that $\lim_{n\to\infty} ||X_n - X|| = 0$, if and only if both conditions $\lim_{n\to\infty} E[X_n Y] = E[XY]$ for all random variables Y satisfying $E[Y^2] < \infty$, and $\lim_{n\to\infty} ||X_n|| = ||X||$ hold.

3. Suppose

$$W_n = \sum_{j=1}^{q} \sigma_j \sqrt{2} \cos(\lambda_j n - V_j),$$

where σ_j, λ_j are positive constants, $j = 1, \ldots, q$, and $V_1, \ldots, V_q$ are independent, uniformly distributed in the interval $(0, 2\pi)$. Show that $\{W_n\}$ is covariance stationary and compute the covariance function.

4. Let $\rho(v) = R(v)/R(0)$ be the correlation function of a covariance stationary process $\{X_n\}$, where

$$X_{n+1} = a_1 X_n + a_2 X_{n-1} + \xi_{n+1},$$

for constants a_1, a_2 and zero mean uncorrelated random variables $\{\xi_n\}$, for which $E[\xi_n^2] = \sigma^2$ and $E[\xi_n X_{n-k}] = 0$, $k = 1, 2, \ldots$. Establish that $\rho(v)$ satisfies the so-called *Yule–Walker* equations

$$\rho(1) = a_1 + a_2 \rho(1), \qquad \text{and} \qquad \rho(2) = a_1 \rho(1) + a_2 .$$

Determine a_1 and a_2 in terms of $\rho(1)$ and $\rho(2)$.

5. Show that no covariance stationary process $\{X_n\}$ can satisfy the stochastic difference equation $X_n = X_{n-1} + \varepsilon_n$, when $\{\varepsilon_n\}$ is a sequence of zero-mean uncorrelated random variables having a common positive variance $\sigma^2 > 0$.

6. Let $\{X_n\}_{n=-\infty}^{+\infty}$ be a zero-mean covariance stationary process having covariance function $R(v) = \gamma^{|v|}$, $v = 0, \pm 1, \ldots$, where $|\gamma| < 1$. Find the minimum mean square error linear predictor of X_{n+1} given the entire past $X_n, X_{n-1} \ldots$.

7. Let $\{\varepsilon_n\}$ be zero-mean uncorrelated random variables having unit variances. Find the minimum mean square linear predictor for X_{n+1} given the entire past $X_n, X_{n-1}, \ldots$, for the moving average process

$$X_n = \varepsilon_n + \beta[\varepsilon_{n-1} + \gamma \varepsilon_{n-2} + \gamma^2 \varepsilon_{n-3} + \cdots],$$

where β and γ are constants, $|\gamma| < 1$, and $|\alpha| < 1$, where $\alpha = \gamma - \beta$.

8. Let $\{X_n\}$ be a zero-mean covariance stationary process having covariance function $R_X(v)$ and spectral density function $f_X(\omega)$, $-\pi \leq \omega \leq \pi$. Suppose $\{a_n\}$ is a real sequence for which $\sum_{i,j=0}^{\infty} |a_i a_j R(i-j)| < \infty$, and define

$$Y_n = \sum_{k=0}^{\infty} a_k X_{n-k}.$$

Show that the spectral density function $f_Y(\omega)$ for $\{Y_n\}$ is given by

$$f_Y(\omega) = \frac{\sigma_X^2}{\sigma_Y^2} \left| \sum_{k=0}^{\infty} a_k e^{ik\omega} \right|^2 f_X(\omega),$$

$$= \frac{\sigma_X^2}{\sigma_Y^2} \left[\sum_{j,k=0}^{\infty} a_j a_k \cos(j-k)\omega \right] f_X(\omega), \qquad -\pi \leq \omega \leq \pi.$$

9. Determine the spectral density function corresponding to the covariance function $R(v) = \gamma^{|v|}$, $v = 0, \pm 1, \ldots$, where $|\gamma| < 1$.

Answer:

$$f(\omega) = \frac{1-\gamma^2}{2\pi|1-\gamma e^{i\omega}|^2}, \qquad -\pi < \omega < \pi.$$

10. Compute the spectral density function of the autoregressive process $\{X_n\}$ satisfying

$$X_n = \beta_1 X_{n-1} + \cdots + \beta_q X_{n-q} + \xi_n,$$

where $\{\xi_n\}$ are uncorrelated zero-mean random variables having unit variance. Assume the q roots of $x^q - \beta_1 x^{q-1} \cdots - \beta_q = 0$ are all less than one in absolute value.

Answer:

$$f(\omega) = \left\{ 2\pi\sigma_X^2 \left| 1 - \sum_{k=1}^{q} \beta_k e^{ik\omega} \right|^2 \right\}^{-1}, \quad -\pi < \omega < \pi.$$

11. Compute the spectral density function of the moving average process

$$X_n = \xi_n + \alpha_1 \xi_{n-1}.$$

Answer:

$$f(\lambda) = \frac{1 + \alpha_1^2 + 2\alpha_1 \cos\lambda}{2\pi(1+\alpha_1^2)},$$

where $\{\xi_n\}$ are uncorrelated zero-mean random variables having unit variance.

12. Let $\{X_n\}$ be the finite moving average process

$$X_n = \sum_{r=0}^{q} \alpha_r \xi_{n-r}, \qquad \alpha_0 = 1,$$

where $\alpha_0, \ldots, \alpha_q$ are real and $\{\xi_n\}$ are zero-mean uncorrelated random variables having unit variance. Show that the spectral density function $f(\lambda)$ may be written

$$f(\lambda) = \frac{1}{2\pi\sigma_X^2} \prod_{j=1}^{q} |e^{i\lambda} - z_j|^2,$$

where $z_1, \ldots, z_q$ are the q roots of

$$\sum_{r=0}^{q} a_r z^{q-r} = 0.$$

13. Show that a predictor

$$\hat{X}_n = \alpha_1 X_{n-1} + \cdots + \alpha_p X_{n-p}$$

is optimal among all linear predictors of X_n given $X_{n-1}, \ldots, X_{n-p}$ if and only if

$$0 = \int_{-\pi}^{\pi} e^{ik\lambda} \left[1 - \sum_{l=1}^{p} \alpha_l e^{-il\lambda}\right] dF(\lambda), \qquad k = 1, \ldots, p,$$

where $F(\omega)$, $-\pi \leq \omega \leq \pi$, is the spectral distribution function of the covariance stationary process $\{X_n\}$.

14. Let $\{X_n\}$ be a zero-mean covariance stationary process having positive spectral density function $f(\omega)$ and variance $\sigma_X^2 = 1$. Kolmogorov's formula states

$$\sigma_e^2 = \exp\left\{\frac{1}{2\pi} \int_{-\pi}^{\pi} \log 2\pi f(\omega)\, d\omega\right\},$$

where $\sigma_e^2 = \inf E[|\hat{X}_n - X_n|^2]$ is the minimum mean square linear prediction error of X_n given the past. Verify Kolmogorov's formula when

$$R(v) = \gamma^{|v|}, \qquad v = 0, \pm 1, \ldots,$$

with $|\gamma| < 1$.

15. Derive

$$R(v) = \int_{-\pi}^{\pi} (\cos \lambda v) f(\lambda)\, d\lambda,$$

from

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} R(k) \cos \lambda k, \qquad -\pi \leq \lambda \leq \pi.$$

16. Compute the covariance function and spectral density function for the moving average process

$$X_n = \sum_{k=0}^{\infty} a_k \xi_{n-k},$$

where $\{\xi_n\}$ are zero-mean uncorrelated random variables having unit variance and $a_0, a_1, \ldots$ are real numbers satisfying $\sum a_k^2 < \infty$.

17. Find the minimum mean square error linear predictor of X_{n+1} given X_n, $X_{n-1}, \ldots, X_0$ in the following nonstationary linear model: $\theta_0, \zeta_1, \zeta_2, \ldots$, and $\varepsilon_0, \varepsilon_1, \ldots$ are all uncorrelated with zero means. The variances are $E[\theta_0^2] = v_0^2$, $E[\zeta_k^2] = v^2$, and $E[\varepsilon_k^2] = \sigma^2$, where $v^2 = \alpha v_0^2$, $\alpha = v_0^2/(v_0^2 + \sigma^2)$. Finally, $X_n = \theta_n + \varepsilon_n$, where $\theta_{n+1} = \theta_n + \zeta_{n+1}$, $n = 0, 1, \ldots$. (We interpret $\{X_n\}$ as a noise distorted observation on the θ process.)

Answer:

$$\hat{X}_0 = 0$$
$$\hat{X}_k = \alpha X_{k-1} + (1 - \alpha)\, X_{k-1}, \text{ for } k = 1, 2, \ldots,$$

where $\alpha = v_0^2/(v_0^2 + \sigma^2)$.

18. Let $\{X_k\}$ be a moving average process

$$X_n = \sum_{j=0}^{\infty} \alpha_j \xi_{n-j}, \qquad \alpha_0 = 1, \qquad \sum_{j=0}^{\infty} \alpha_j^2 < \infty,$$

where $\{\xi_n\}$ are zero-mean independent random variables having common variance σ^2. Show that

$$U_n = \sum_{k=0}^{n} X_{k-1}\xi_k, \qquad n = 0, 1, \ldots,$$

and

$$V_n = \sum_{k=0}^{n} X_k \xi_k - (n+1)\sigma^2, \qquad n = 0, 1, \ldots,$$

are martingales with respect to $\{\xi_n\}$.

19. Let $i = \sqrt{-1}$. Define the integral of a complex-valued function with respect to Gaussian random measure as the sum of the integrals of the real and imaginary parts. Similarly, define the integral of a function with respect to a complex random measure

$$\zeta(\mathrm{I}) = \xi(\mathrm{I}) + i\eta(\mathrm{I}), \qquad \mathrm{I} = (s, t], \qquad s < t,$$

as the sum of the real and imaginary parts. Obtain the representation

$$X_n = \int_{-\pi}^{\pi} e^{-in\omega} \zeta\,(d\omega), \qquad n = 0, \pm 1, \ldots,$$

from the representation

$$X_n = \int_0^{\pi} \cos n\omega Z^{(1)}\,(d\omega) + \int_0^{\pi} \sin n\omega Z^{(2)}\,(d\omega),$$

by setting

$$\xi(s) = -\xi(-s) = \tfrac{1}{2}Z^{(1)}(s), \qquad 0 \leq s \leq \pi,$$

and

$$\eta(s) = \eta(-s) = \tfrac{1}{2}Z^{(2)}(s), \qquad 0 \leq s \leq \pi.$$

Observe that $\xi(I) = \xi(-I)$ and $\eta(I) = -\eta(-I)$, where $I = (s, t]$, $-I = (-t, -s]$, $0 \leq s \leq t \leq \pi$. Compute $E[\zeta(I_1)\overline{\zeta(I_2)}]$, where $I_i = (s_i, t_i]$ and "bar" denotes complex conjugation.

20. Suppose X_0 has probability density function

$$f(x) = \begin{cases} 2x, & \text{for } 0 \leq x \leq 1, \\ 0, & \text{elsewhere,} \end{cases}$$

and that X_{n+1} is uniformly distributed on $(1 - X_n, 1]$, given $X_0, \ldots, X_n$. Show that $\{X_n\}$ is a stationary ergodic process.

21. Show that every sequence $X_1, X_2, \ldots$ of independent identically distributed random variables forms an ergodic stationary process.

22. Let Z be a random variable uniformly distributed on $[0, 1)$. Let $X_0 = Z$ and $X_{n+1} = 2X_n \pmod 1$ that is,

$$X_{n+1} = \begin{cases} 2X_n & \text{if } X_n < \frac{1}{2}, \\ 2X_n - 1 & \text{if } X_n \geq \frac{1}{2}. \end{cases}$$

(a) Show that if $Z = .Z_0 Z_1 Z_2 \ldots$ is the terminating binary expansion for $Z = \sum_{k=0}^{\infty} 2^{-(k+1)} Z_k$, then $X_n = .Z_n Z_{n+1} \ldots$. (b) Show that X_n is a stationary process.

(c) Show that $\{X_n\}$ is ergodic. (d) Use the ergodic theorem to show that with probability one

$$\frac{1}{n} \sum_{k=0}^{n-1} \{2^k Z\} \to \frac{1}{2},$$

where $\{x\}$ is the fractional part of x ($\{x\} = x - [x]$ with $[x]$ the largest integer not exceeding x).

23. Let $\{\xi_n\}$ be independent identically distributed random variables having zero means and unit variances. Show that every moving average

$$X_n = \sum_{k=0}^{m} a_k \xi_{n-k}, \qquad n = 0, \pm 1, \ldots,$$

is ergodic. Suppose $\sum a_k^2 < \infty$. Is the same true of

$$Y_n = \sum_{k=0}^{\infty} a_k \xi_{n-k}?$$

24. A stochastic process $\{X_n\}$ is said to be *weakly mixing* if, for all sets A, B of real sequences $(x_1, x_2, \ldots)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \Pr\{(X_1, X_2, \ldots) \in A \quad \text{and} \quad (X_k, X_{k+1}, \ldots) \in B\}$$
$$= \Pr\{(X_1, X_2, \ldots) \in A\} \times \Pr\{(X_1, X_2, \ldots) \in B\}.$$

Show that every weakly mixing process is ergodic.

Remark: To verify weakly mixing, it suffices to show, for every $m = 1, 2, \ldots$, and all sets A,B of vectors $(x_1, \ldots, x_m)$, that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \Pr\{(X_1, \ldots, X_m) \in A \quad \text{and} \quad (X_{k+1}, \ldots, X_{k+m}) \in B\}$$
$$= \Pr\{(X_1, \ldots, X_m) \in A\} \times \Pr\{(X_1, \ldots, X_m) \in B\}.$$

25. Let $\{\xi_n\}$ be a zero-mean covariance stationary process having covariance function

$$E[\xi_n \xi_m] = \begin{cases} 1, & n = m, \\ \rho, & n \neq m, \end{cases}$$

where $0 < \rho < 1$. Show that $\{\xi_n\}$ has the representation $\xi_n = U + \eta_n$, where $U, \eta_1, \eta_2, \ldots$ are zero-mean, uncorrelated random variables, $E[U^2] = \rho$, and $E[\eta_k^2] = 1 - \rho$.

Hint: Use the mean square ergodic theorem to define $U = \lim(\xi_1 + \cdots + \xi_n)/n$. Set $\eta_n = \xi_n - U$ and compute $E[U\xi_n]$, $E[U^2]$, and $E[\eta_n \eta_m]$.

26. Let $\{X_n\}$ be a finite-state irreducible Markov chain having the transition probabilities $||P_{ij}||_{i,j=1}^{N}$. There then exists a stationary distribution π, i.e., a vector $\pi(1), \ldots, \pi(N)$ satisfying $\pi(i) \geq 0$, $i = 1, \ldots, N$, $\sum_{i=1}^{N} \pi(i) = 1$, and

$$\pi(j) = \sum_{i=1}^{N} \pi(i) P_{ij}, \qquad j = 1, \ldots, N.$$

Suppose $\Pr\{X_0 = i\} = \pi(i)$, $i = 1, \ldots, N$. Show that $\{X_n\}$ is weakly mixing, hence ergodic.

27. Let $\{B(t), t \geq 0\}$ be a standard Brownian motion process and $B(I) = B(t) - B(s)$, for $I = (s, t]$, $0 \leq s < t$ the associated Gaussian random measure. Show that

$$Y(t) = \int_0^t f(x)B(dx), \qquad t \geq 0,$$

is a martingale for every bounded continuous function $f(u)$, $u \geq 0$.

28. Let $\{B(t), t \geq 0\}$ be a standard Brownian motion process and $B(I) = B(t) - B(s)$, for $I = (s, t]$, $0 \leq s < t$ the associated Gaussian random measure. Let $f(u)$ be a continuous function for $u \in [0, h]$. Show

$$Y(t) = \int_t^{t+h} f(u-t)B(du), \qquad t \geq 0,$$

is a stationary process.

29. Under the conditions of Theorem 10.2, show

$$\lim_{\omega \to \infty} \Pr\left\{ \max_{0 \leq s \leq t/f(\omega)} X(s) < \omega \right\} = e^{-t},$$

where

$$f(\omega) = \frac{\sqrt{\lambda_2}}{2\pi\sigma} \exp(-\omega^2/2\sigma^2).$$

30. Let $\{B(t);\ 0 \leq t \leq 1\}$ be a standard Brownian motion process and let $B(I) = B(t) - B(s)$, for $I = (s, t]$, $0 \leq s \leq t \leq 1$ be the associated Gaussian random measure. Validate the identity

$$E\left[\exp\left\{\lambda \int_0^1 f(s)\, dB(s)\right\}\right] = \exp\left\{\tfrac{1}{2}\lambda^2 \int_0^1 f^2(s)\, ds\right\}, \quad -\infty < \lambda < \infty$$

where $f(s)$, $0 \leq s \leq 1$ is a continuous function.

31. Let $\{B(t);\ 0 \leq t \leq 1\}$ be a standard Brownian motion process and let $B(I) = B(t) - B(s)$, for $I = (s, t]$, $0 \leq s \leq t \leq 1$ be the associated Gaussian random measure. Validate the assertion that $U = \int_0^1 f(s)\, dB(s)$ and $V = \int_0^1 g(s)\, dB(s)$ are independent random variables whenever f and g are bounded continuous functions satisfying $\int_0^1 f(s)\, g(s)\, ds = 0$.

NOTES

For a good introduction to the spectral theory of stationary processes, see the book by Yaglom [2].

Many aspects of stationary processes, including the level-crossing problem, are treated in the text by Cramer and Leadbetter [1].

REFERENCES

1. Cramer, H, and Leadbetter, M., "Stationary and Related Stochastic Processes." Wiley, New York, 1966.
2. Yaglom, A. M., "An Introduction to the Theory of Stationary Random Functions." Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
3. Anderson, T. W., "Statistical Analysis of Time Series." Wiley, New York, 1973.

Appendix

REVIEW OF MATRIX ANALYSIS

1: The Spectral Theorem

A. INTRODUCTORY CONCEPTS (LINEAR INDEPENDENCE AND BASIS)†

The set of all n-tuples (vectors) $\mathbf{x} = (x_1, \ldots, x_n)$, where the x_i are complex numbers, forms what is called an n-dimensional vector space. The sum of two vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ is defined by $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n)$, and the product of $\mathbf{x}$ by a complex number λ is defined by $\lambda\mathbf{x} = (\lambda x_1, \ldots, \lambda x_n)$.

A set $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(r)}$ of vectors is called *linearly independent* if the relation

$$c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)} + \cdots + c_r\mathbf{x}^{(r)} = \mathbf{0}$$

implies $c_1 = c_2 = \cdots = c_r = 0$. As an example, we exhibit the vectors $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, etc., whose linear independence is obvious. A linearly independent family of vectors of n-tuples cannot contain more than n vectors.

Let $\varphi_1, \ldots, \varphi_r$, $r < n$, be a linearly independent set. There is some vector φ_{r+1} which is not a linear combination of $\varphi_1, \ldots, \varphi_r$, i.e., of the form $c_1\varphi_1 + \cdots + c_r\varphi_r$. It follows at once that $\varphi_1, \ldots, \varphi_{r+1}$ is a linearly independent set. Proceeding in this fashion, we may obviously augment

† Some statements are made without formal proofs; the industrious student should supply detailed arguments.

the set $\varphi_1, \ldots, \varphi_r$ by vectors $\varphi_{r+1}, \ldots, \varphi_n$ so that $\varphi_1, \ldots, \varphi_n$ is a linearly independent set. Since no linearly independent set can contain more than n elements, we can determine for each vector $\mathbf{y}$ and each linearly independent set $\varphi_1, \ldots, \varphi_n$ constants $c_1, \ldots, c_n$ (necessarily unique) such that $c_1\varphi_1 + \cdots + c_n\varphi_n = \mathbf{y}$.

Analogous results hold for any linear manifold $\mathfrak{W}$, i.e., any set $\mathfrak{W}$ of vectors such that $\mathbf{x}, \mathbf{y} \in \mathfrak{W}$ implies $a\mathbf{x} + b\mathbf{y} \in \mathfrak{W}$ for every complex a, b. Given a linear manifold $\mathfrak{W}$, there is a unique integer m, $0 \leq m \leq n$, the dimension of $\mathfrak{W}$, such that the largest linearly independent set in $\mathfrak{W}$ contains precisely m elements. If $\varphi_1, \ldots, \varphi_r$, $r < m$, is a linearly independent set lying in $\mathfrak{W}$, there is some vector $\varphi_{r+1} \in \mathfrak{W}$ which is not expressible as a linear combination of $\varphi_1, \ldots, \varphi_r$. As before we infer the existence of $\varphi_{r+1}, \ldots, \varphi_m \in \mathfrak{W}$ such that $\varphi_1, \ldots, \varphi_m$ is a linearly independent set. Moreover, for any $\mathbf{y} \in \mathfrak{W}$ there exist (necessarily unique) constants $c_1, \ldots, c_m$ for which $c_1\varphi_1 + \cdots + c_m\varphi_m = \mathbf{y}$ holds. Notice that $\dim \mathfrak{W} = 0$ implies that $\mathfrak{W}$ consists exclusively of the zero element, while $\dim \mathfrak{W} = n$ implies that it contains every vector. If $\dim \mathfrak{W} = m$, any set of m linearly independent vectors in $\mathfrak{W}$ is called a ***basis*** of $\mathfrak{W}$. The unqualified term "basis" will be used for any set of n linearly independent vectors.

B. SCALAR PRODUCTS

The scalar (also called inner) product of two vectors $\mathbf{x}, \mathbf{y}$ is defined by $(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i\bar{y}_i$, where $\bar{y}_i$ denotes the complex conjugate of y_i. We note the following easily verified properties of the scalar product:

(i) $(\mathbf{x}, \mathbf{x}) \geq 0$, with equality if and only if $\mathbf{x} = (0, \ldots, 0) = \mathbf{0}$.
(ii) $(\lambda\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{x}, \mathbf{y})$ for λ complex.
(iii) $(\mathbf{x}, \mathbf{y}) = \overline{(\mathbf{y}, \mathbf{x})}$; thus from (ii), $(\mathbf{x}, \lambda\mathbf{y}) = \bar{\lambda}(\mathbf{x}, \mathbf{y})$.

Two vectors $\mathbf{x}, \mathbf{y}$ are called orthogonal if $(\mathbf{x}, \mathbf{y}) = 0$. The norm $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ is defined by $\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{\frac{1}{2}}$.

A set $\{a_{ij}\}$, $i, j = 1, \ldots, n$, of complex numbers defines an n-dimensional (square) matrix, usually denoted by $\mathbf{A} = \|a_{ij}\|$, $i, j = 1, \ldots, n$. An $n \times n$ matrix $\mathbf{A}$ defines a transformation (or operator) in an n-dimensional vector space according to either $\mathbf{Ax} = \mathbf{y}$, where $y_i = \sum_{j=1}^{n} a_{ij}x_j$, $i = 1, \ldots, n$, or $\mathbf{xA} = \mathbf{z}$, where $z_j = \sum_{i=1}^{n} x_i a_{ij}$, $j = 1, \ldots, n$. It is immediate from these definitions that

$$\mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{Ax} + \beta\mathbf{Ay}, \qquad (\alpha\mathbf{x} + \beta\mathbf{y})\mathbf{A} = \alpha\mathbf{xA} + \beta\mathbf{yA},$$

for any vectors $\mathbf{x}, \mathbf{y}$ and constants α, β. Further, $(\mathbf{x}, \mathbf{Ay}) = (\mathbf{x}\bar{\mathbf{A}}, \mathbf{y})$ where $\bar{\mathbf{A}}$ is the $n \times n$ matrix with elements $\bar{a}_{ij}$. The two transformations induced

by operating on the right or left are appropriately dual to each other. An elaborate geometric and algebraic theory of linear transformations is available; it can be found in most textbooks on matrix theory.

C. EIGENVALUES AND EIGENVECTORS

A complex number λ is called an eigenvalue of the matrix $\mathbf{A}$ if there exists a vector $\mathbf{x}^{(\lambda)} \neq 0$ such that $\mathbf{A}\mathbf{x}^{(\lambda)} = \lambda\mathbf{x}^{(\lambda)}$. If λ is an eigenvalue of $\mathbf{A}$, then the set $\mathfrak{W}_\lambda$ consisting of all vectors which satisfy the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is called the right eigenmanifold of $\mathbf{A}$ corresponding to the eigenvalue λ, and the members of $\mathfrak{W}_\lambda$ are called right eigenvectors for λ. Clearly $\mathbf{y}$ and $\mathbf{z} \in \mathfrak{W}_\lambda$ implies $a\mathbf{y} + b\mathbf{z} \in \mathfrak{W}_\lambda$ for any constants a, b. The dimension of $\mathfrak{W}_\lambda$ is called the geometric multiplicity of λ.

If $\mathfrak{W}_1, \ldots, \mathfrak{W}_r$ are distinct eigenmanifolds of the operator $\mathbf{A}$, and $\varphi_1, \ldots, \varphi_r$ are arbitrary nonzero vectors in $\mathfrak{W}_1, \ldots, \mathfrak{W}_r$, respectively, then $\varphi_1, \ldots, \varphi_r$ are linearly independent. In fact, supposing the contrary, we let m be the smallest integer for which we can find distinct $\mathfrak{W}_1, \ldots, \mathfrak{W}_m$ and associated nonzero vectors $\varphi_1 \in \mathfrak{W}_1, \ldots, \varphi_m \in \mathfrak{W}_m$, with accompanying nonzero constants $c_1, \ldots, c_m$ such that $c_1\varphi_1 + \cdots + c_m\varphi_m = 0$. Trivially, $m \geq 2$. Applying $\mathbf{A}$ to both sides of the last equation, we obtain $\lambda_1 c_1\varphi_1 + \cdots + \lambda_m c_m\varphi_m = \mathbf{0}$, where λ_i is the eigenvalue corresponding to the eigenmanifold $\mathfrak{W}_i$. If one among the λ_i is zero, we have a linear relationship among $m - 1$ elements, which contradicts the definition of m. Thus, $\lambda_1 \neq 0$; multiplying $c_1\varphi_1 + \cdots + c_m\varphi_m = \mathbf{0}$ by λ_1 and subtracting the result from $\lambda_1 c_1\varphi_1 + \cdots + \lambda_m c_m\varphi_m = \mathbf{0}$, we have

$$(\lambda_2 - \lambda_1)c_2\varphi_2 + \cdots + (\lambda_m - \lambda_1)c_m\varphi_m = \mathbf{0}.$$

This again contradicts the definition of m. It follows at once that if $\mathfrak{W}_1, \ldots, \mathfrak{W}_r$ are distinct eigenmanifolds of $\mathbf{A}$, and $\varphi_1^{(i)}, \ldots, \varphi_{m_i}^{(i)}$ is a basis of $\mathfrak{W}_i$, $i = 1, \ldots, r$, then

$$\varphi_1^{(1)}, \ldots, \varphi_{m_1}^{(1)}, \quad \varphi_1^{(2)}, \ldots, \varphi_{m_2}^{(2)}, \ldots, \quad \varphi_1^{(r)}, \ldots, \varphi_{m_r}^{(r)}$$

form a linearly independent set. Therefore, $\mathbf{A}$ can have only a finite number of eigenvalues and eigenmanifolds. In the important case where the sum of the dimensions of the eigenmanifolds equals n, we can construct a basis (for the entire space) composed of eigenvectors only. A matrix with this property is called *diagonalizable*.

We could just as well have started with the equation $\mathbf{x}\mathbf{A} = \lambda\mathbf{x}$ in place of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. It turns out that the values of λ for which $\mathbf{x}\mathbf{A} = \lambda\mathbf{x}$ has a nontrivial solution are precisely the eigenvalues of $\mathbf{A}$ as defined in the preceding paragraph. Furthermore, the dimension of the manifold of vectors satisfying $\mathbf{x}\mathbf{A} = \lambda\mathbf{x}$ (i.e., *left* eigenvectors) is just the multiplicity

of λ. (The reader should prove this fact.) As before, a set consisting of left eigenvectors, each associated with a different eigenvalue, is linearly independent.

It may be noted that the eigenvalues of $\mathbf{A}$ are precisely the roots of the nth-degree algebraic equation

$$\det\|\mathbf{A} - \lambda\mathbf{I}\| = 0.$$

From this and the known properties of determinants follows a result which we will need later on, namely, that if

$$\mathbf{A} = \begin{Vmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{B} & \mathbf{A}_2 \end{Vmatrix},$$

where $\mathbf{A}_1$, $\mathbf{A}_2$ are square matrices, then a number λ is an eigenvalue of $\mathbf{A}$ if and only if it is an eigenvalue of (at least) one of the matrices $\mathbf{A}_1$, $\mathbf{A}_2$. In fact, since

$$\det\|\mathbf{A} - \lambda\mathbf{I}\| = \det\|\mathbf{A}_1 - \lambda\mathbf{I}\| \det\|\mathbf{A}_2 - \lambda\mathbf{I}\|,$$

where the same notation $\mathbf{I}$ is used for identity matrices of varying dimensions, the assertion is obvious.

(a) *Spectral Representation*

For the following discussion we assume that $\mathbf{A}$ is real, i.e., its elements are real. Suppose that we can construct a basis for the whole space using right eigenvectors of $\mathbf{A}$. From the preceding remarks it follows that we can construct a basis for the whole space, using left eigenvectors of $\mathbf{A}$ as well. If in addition the elements a_{ij} of $\mathbf{A}$ are all real, we can choose the two bases to be biorthogonal, i.e., $\varphi_1, \ldots, \varphi_n$ and $\psi_1, \ldots, \psi_n$ are respectively the bases consisting of right and left eigenvectors for which $(\varphi_i, \psi_j) = 1$ if $i = j$ and 0 otherwise. To demonstrate the construction of eigenvectors with these properties, we note first that if $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$ and $\mathbf{y}_j \mathbf{A} = \mu_j \mathbf{y}_j$, then

$$\mu_j(\mathbf{y}_j, \mathbf{x}_i) = (\mu_j \mathbf{y}_j, \mathbf{x}_i) = (\mathbf{y}_j \mathbf{A}, \mathbf{x}_i) = (\mathbf{y}_j, \mathbf{A}\mathbf{x}_i) = (\mathbf{y}_j, \lambda_i \mathbf{x}_i) = \bar{\lambda}_i(\mathbf{y}_j, \mathbf{x}_i),$$

so that if $\mu_j \neq \bar{\lambda}_i$ we must have $(\mathbf{y}_j, \mathbf{x}_i) = 0$. Next we observe that, because $\mathbf{A}$ is real, it follows directly that when $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ holds, $\mathbf{A}\bar{\mathbf{x}} = \bar{\lambda}\bar{\mathbf{x}}$ also holds, where $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n)$. We see, therefore, that the eigenvalues of $\mathbf{A}$ occur in conjugate pairs, and λ and $\bar{\lambda}$ have the same multiplicity. Let the eigenvalues of $\mathbf{A}$ be

$$\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2, \ldots, \lambda_r, \bar{\lambda}_r, \lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_m,$$

where $\lambda_1, \ldots, \lambda_r$ are complex and $\lambda_{r+1}, \ldots, \lambda_m$ are real. We denote the corresponding right eigenmanifolds by $\mathfrak{M}_1, \overline{\mathfrak{M}}_1, \ldots, \mathfrak{M}_r, \overline{\mathfrak{M}}_r, \mathfrak{M}_{r+1}, \ldots, \mathfrak{M}_m$ and the left eigenmanifolds by $\mathfrak{L}_1, \overline{\mathfrak{L}}_1, \ldots, \mathfrak{L}_r, \overline{\mathfrak{L}}_r, \mathfrak{L}_{r+1}, \ldots, \mathfrak{L}_m$.

Now we have shown that every element in $\mathfrak{L}_1$ is orthogonal to every element in all of the right eigenmanifolds except $\overline{\mathfrak{W}}_1$, and similarly for the other left eigenmanifolds. Our task is, therefore, reduced to selecting a basis $\psi_1, \ldots, \psi_d$ for $\mathfrak{L}_1$ and $\varphi_1, \ldots, \varphi_d$ for $\overline{\mathfrak{W}}_1$ such that $(\psi_i, \varphi_j) = 1$ if $i = j$ and 0 otherwise, where d is the multiplicity of λ_1, and similarly for $\overline{\mathfrak{L}}_1, \overline{\mathfrak{W}}_1, \mathfrak{L}_2, \overline{\mathfrak{W}}_2$. To do this, let $\varphi_1, \ldots, \varphi_d$ be any basis for $\mathfrak{W}_1$, and $\mathbf{y}_1, \ldots, \mathbf{y}_d$ any basis for $\mathfrak{L}_1$. We wish to choose constants $c_1, \ldots, c_d$ such that $\psi_1 = c_1 \mathbf{y}_1 + \cdots + c_d \mathbf{y}_d$, and fullfilling the conditions $(\psi_1, \varphi_1) = 1$ and $(\psi_1, \varphi_i) = 0$ for $i = 2, \ldots, d$; i.e., we wish to satisfy the relations

$$\begin{aligned}
&c_1(\mathbf{y}_1, \varphi_1) + c_2(\mathbf{y}_2, \varphi_1) + \cdots + c_d(\mathbf{y}_d, \varphi_1) = 1,\\
&c_1(\mathbf{y}_1, \varphi_2) + c_2(\mathbf{y}_2, \varphi_2) + \cdots + c_d(\mathbf{y}_d, \varphi_2) = 0,\\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad\quad \vdots\\
&c_1(\mathbf{y}_1, \varphi_d) + c_2(\mathbf{y}_2, \varphi_d) + \cdots + c_d(\mathbf{y}_d, \varphi_d) = 0.
\end{aligned}$$

Suppose that this system of linear equations for $c_1, \ldots, c_d$ has no solution. This means that no linear combination of the d vectors

$$\mathbf{f}_1 = ((\mathbf{y}_1, \varphi_1), \ldots, (\mathbf{y}_1, \varphi_d)), \ldots, \mathbf{f}_d = ((\mathbf{y}_d, \varphi_1), \ldots, (\mathbf{y}_d, \varphi_d))$$

yields the vector $(1, 0, \ldots, 0)$, and therefore the vectors $\mathbf{f}_1, \ldots, \mathbf{f}_d$ are not linearly independent. Thus, there exist constants $a_1, \ldots, a_d$, not all zero, such that $a_1\mathbf{f}_1 + \cdots + a_d\mathbf{f}_d = \mathbf{0}$. But this implies the equation

$$(a_1 \mathbf{y}_1 + \cdots + a_d \mathbf{y}_d, \varphi_i) = 0, \qquad i = 1, \ldots, d.$$

But it was proved above that the set $\mathbf{y}_1, \ldots \mathbf{y}_d$ (and hence any linear combination of the $\mathbf{y}_i$) is orthogonal to every manifold of right eigenvectors except $\overline{\mathfrak{W}}_1$. Now we see that $a_1 \mathbf{y}_1 + \cdots + a_d \mathbf{y}_d$ is orthogonal to every right eigenvector, and, of course, every linear combination of right eigenvectors. But by assumption there exists a basis formed of right eigenvectors, so that $a_1\mathbf{y}_1 + \cdots + a_d\mathbf{y}_d$ is orthogonal to itself, and hence equals $\mathbf{0}$. This contradicts the linear independence of $\mathbf{y}_1, \ldots, \mathbf{y}_d$. Thus ψ_1 having the desired properties exists. We construct $\psi_2, \ldots, \psi_d$ in a similar manner. It remains to show that $\psi_1, \ldots, \psi_d$ are linearly independent. Suppose that $a_1\psi_1 + \cdots + a_d\psi_d = 0$; then

$$\begin{aligned}
0 &= (a_1\psi_1 + \cdots + a_d\psi_d, \varphi_1) = a_1,\\
0 &= (a_1\psi_1 + \cdots + a_d\psi_d, \varphi_2) = a_2,\\
&\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad \vdots\\
0 &= (a_1\psi_1 + \cdots + a_d\psi_d, \varphi_d) = a_d,
\end{aligned}$$

and the above implication establishes that the ψ_i are linearly independent.

As we have seen, if $\mathbf{A}$ is a matrix whose elements are all real, and whose right (and therefore left) eigenvectors can be chosen to be a basis for the whole space, we can indeed take the basis $\varphi_1, \ldots, \varphi_n$ of right eigenvectors and $\psi_1, \ldots, \psi_n$ of left eigenvectors to be mutually biorthogonal, i.e., satisfying the relations $(\psi_i, \varphi_j) = 1$ if $i = j$, and 0 otherwise.

We will make use of this result to develop a canonical representation of the matrix $\mathbf{A}$, the so-called spectral representation. Suppose that $\lambda_1, \ldots, \lambda_n$ are the eigenvalues corresponding to $\varphi_1, \ldots, \varphi_n$, respectively; i.e., $\mathbf{A}\varphi_i = \lambda_i\varphi_i$, $i = 1, \ldots, n$, where the λ_i need not be distinct. Let

$$\varphi_i = (\varphi_{i1}, \ldots, \varphi_{in}), \qquad \psi_i = (\psi_{i1}, \ldots, \psi_{in}),$$

$$\mathbf{\Phi} = \begin{Vmatrix} \varphi_{11} \cdots \varphi_{n1} \\ \vdots \\ \varphi_{1n} \cdots \varphi_{nn} \end{Vmatrix}, \qquad \mathbf{\Psi} = \begin{Vmatrix} \psi_{11} \cdots \psi_{1n} \\ \vdots \\ \psi_{n1} \cdots \psi_{nn} \end{Vmatrix},$$

$$\mathbf{\Lambda} = \begin{Vmatrix} \lambda_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2 & & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & & \cdot & \\ 0 & 0 & \cdot & \cdot & \cdot & \lambda_n \end{Vmatrix}.$$

The biorthogonality of $\varphi_1, \ldots, \varphi_n$ and $\psi_1, \ldots, \psi_n$ shows at once that $\mathbf{\Psi\Phi} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Moreover, by direct computation we readily verify that $\mathbf{\Phi\Lambda\Psi}\varphi_i = \lambda_i\varphi_i$, $i = 1, \ldots, n$. Since $\mathbf{A}\varphi_i = \lambda_i\varphi_i$, $i = 1, \ldots, n$, and the φ_i form a basis for the whole space, it follows that

$$\mathbf{A} = \mathbf{\Phi\Lambda\Psi} \qquad \text{and} \qquad \mathbf{\Phi\Psi} = \mathbf{I}.$$

From this we see that $\mathbf{A}^m = \mathbf{\Phi\Lambda\Psi\Phi\Lambda\Psi} \cdots \mathbf{\Phi\Lambda\Psi} = \mathbf{\Phi\Lambda}^m\mathbf{\Psi}$. But

$$\mathbf{\Lambda}^m = \begin{Vmatrix} \lambda_1^m & 0 \cdots 0 \\ 0 & \lambda_2^m \cdots 0 \\ \cdot & \cdot\ \cdot \\ \cdot & \cdot\ \cdot \\ \cdot & \cdot\ \cdot \\ 0 & 0 \cdots \lambda_n^m \end{Vmatrix},$$

and so $\mathbf{A}^m$ is relatively easy to compute, once its spectral representation is explicitly known.

(b) *Convergence*

We will need to have a notion of convergence for sequences of vectors and sequences of matrices.

A sequence $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ of vectors, in a given (n-dimensional) space, is said to converge to a vector $\mathbf{x}^{(0)}$ if

$$\lim_{j\to\infty} x_i^{(j)} = x_i^{(0)}, \qquad i = 1, \ldots, n.$$

Similarly, a sequence $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots$ of n-dimensional square matrices is said to converge to a matrix $\mathbf{A}^{(0)}$ if

$$\lim_{h\to\infty} a_{ij}^{(h)} = a_{ij}^{(0)}, \qquad i, j = 1, \ldots, n.$$

As an elementary consequence of these definitions, we observe that if $\lim_{h\to\infty} \mathbf{A}^{(h)} = \mathbf{A}^{(0)}$ and $\lim_{j\to\infty} \mathbf{x}^{(j)} = \mathbf{x}^{(0)}$, then $\lim_{h\to\infty} \mathbf{A}^{(h)}\mathbf{x}^{(h)} = \mathbf{A}^{(0)}\mathbf{x}^{(0)}$. Furthermore, if $\{\mathbf{A}^{(j)}\}$ is a sequence of matrices for which there exists a matrix $\mathbf{A}^{(0)}$ and a basis $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ of the whole space, such that

$$\lim_{h\to\infty} \mathbf{A}^{(h)}\mathbf{x}^{(i)} = \mathbf{A}^{(0)}\mathbf{x}^{(i)}, \qquad i = 1, \ldots, n,$$

then $\lim_{h\to\infty} \mathbf{A}^{(h)} = \mathbf{A}^{(0)}$. In fact, it is clear that then $\lim_{h\to\infty} \mathbf{A}^{(h)}\mathbf{y} = \mathbf{A}^{(0)}\mathbf{y}$ whenever $\mathbf{y} = c_1\mathbf{x}^{(1)} + \cdots + c_n\mathbf{x}^{(n)}$, i.e., for every $\mathbf{y}$, since $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ form a basis.

D. POSITIVE DEFINITE MATRICES

An $n \times n$ real matrix $A = \|a_{ij}\|$ is said to be *positive definite* if $\sum_{i,j} a_{ij}x_i x_j > 0$ unless every $x_i = 0$. For the most part we consider only symmetric positive definite matrices, those for which $a_{ij} = a_{ji}$.

A real symmetric positive definite matrix is nonsingular, and all its eigenvalues are strictly positive. Every such matrix has a "square root", in the sense that there exists a real nonsingular matrix $P = \|p_{ij}\|$ for which $A = PP^T$. Here P^T denotes the *transpose* of P, the matrix having elements $p_{ij}^T = p_{ji}$.

2: The Frobenius Theory of Positive Matrices

The Frobenius theory of positive matrices serves usefully in numerous applications of probability theory, particularly in the analysis of Markov transition matrices. We proceed to the development of several aspects of this theory.

Preliminaries

Suppose that $\mathbf{A} = \|a_{ij}\|$, $i, j = 1, \ldots, n$, is a square matrix. If every a_{ij} is

nonnegative, we write $\mathbf{A} \geq \mathbf{0}$; if $\mathbf{A} \geq \mathbf{0}$ and at least one a_{ij} is positive, we write $\mathbf{A} > \mathbf{0}$ and call $\mathbf{A}$ a positive matrix; if every a_{ij} is positive, we write $\mathbf{A} \gg 0$. We use the same notation for a vector $\mathbf{x} = (x_1, \ldots, x_n)$, i.e., $\mathbf{x} \geq \mathbf{0}$ requires that $x_i \geq 0$, $i = 1, \ldots, n$; $\mathbf{x} > \mathbf{0}$ implies that $\mathbf{x} \geq \mathbf{0}$ and $x_i > 0$ for at least one i, and $\mathbf{x} \gg \mathbf{0}$ implies that $x_i > 0$, $i = 1, \ldots, n$. We also write $\mathbf{x} \geq \mathbf{y}$ if $\mathbf{x} - \mathbf{y} \geq \mathbf{0}$, etc. Clearly $\mathbf{A} \geq \mathbf{0}$ and $\mathbf{x} \geq \mathbf{y}$ imply $\mathbf{Ax} \geq \mathbf{Ay}$, while $\mathbf{A} \gg \mathbf{0}$ and $\mathbf{x} > \mathbf{y}$ imply $\mathbf{Ax} \gg \mathbf{Ay}$.

Let $\mathbf{A} \geq \mathbf{0}$, and let Λ be the set consisting of all real numbers λ to each of which corresponds a vector $\mathbf{x} = (x_1, \ldots, x_n)$ such that

$$\sum_{i=1}^{n} x_i = 1, \qquad \mathbf{x} > \mathbf{0}, \qquad \text{and} \qquad \mathbf{Ax} \geq \lambda \mathbf{x}.$$

Let $\lambda_0 = \sup_\Lambda \lambda$; then λ_0 is finite, and it is easy to verify that if $\mathbf{A} \gg \mathbf{0}$ then λ_0 is positive. In fact, if $M = \max_{1 \leq i,j \leq n} a_{ij}$, then $\sum_{i=1}^{n} x_i = 1$ and $\mathbf{x} > \mathbf{0}$ implies that $\sum_{j=1}^{n} a_{ij} x_j \leq M \sum_{j=1}^{n} x_j = M$, $i = 1, \ldots, n$, while $x_j \geq 1/n$ for at least one value of j. It follows that $\lambda_0 \leq nM$. Similarly, if $\mathbf{A} \gg \mathbf{0}$, and $\mathbf{x} > \mathbf{0}$, then $0 < \delta = \min_{1 \leq i,j \leq n} a_{ij}$ and $\sum_{j=1}^{n} a_{ij}(1/n) \geq \delta$, $i = 1, \ldots, n$, from which it follows at once that $\lambda_0 \geq \delta n$.

Suppose that, for a matrix $\mathbf{A} > \mathbf{0}$, we have $\lambda_0 = 0$. If $\mathbf{x} \gg \mathbf{0}$, since $\lambda_0 = 0$ it follows that $\mathbf{Ax}$ cannot be $\gg \mathbf{0}$. Since $\mathbf{Ax} = \mathbf{0}$ for some $\mathbf{x} \gg \mathbf{0}$ evidently requires $\mathbf{A} = \mathbf{0}$. Therefore, there exists some $\mathbf{x} \gg \mathbf{0}$ for which $\mathbf{Ax} > \mathbf{0}$. Let C_1 be the set of indices of the positive components of $\mathbf{Ax}$; obviously C_1 does not depend upon the choice of $\mathbf{x} \gg \mathbf{0}$. Let $\mathbf{y} = (y_1, \ldots, y_n) > 0$ be such that $y_i > 0$ if $i \in C_1$, $y_i = 0$ if $i \notin C_1$, and define C_2 to be the indices of the positive components of $\mathbf{Ay}$. Again C_2 does not depend upon the explicit choice of $\mathbf{y}$, and $C_2 \subseteq C_1$. Since $\lambda_0 = 0$, we may conclude in fact that $C_2 \neq C_1$. Continuing in this manner, we find

$$C_1 \supset C_2 \supset \cdots \supset C_m = C_{m+1} = \cdots = \theta,$$

where the indicated inclusions are all proper. We assert now that $\mathbf{A}^m = \mathbf{0}$. In fact, it is clear that $\mathbf{A}^m \mathbf{x} = \mathbf{0}$ for $\mathbf{x} \gg \mathbf{0}$, and since every vector can be written as the difference of two strictly positive vectors, $\mathbf{A}^m \mathbf{z} = \mathbf{0}$ for every $\mathbf{z}$, which is equivalent to $\mathbf{A}^m = \mathbf{0}$.

The First Frobenius Theorem. We are now in a position to prove the first principal Frobenius theorem.

Theorem 2.1. *If* $\mathbf{A} \gg \mathbf{0}$, *then* (a) *there exists* $\mathbf{x}^0 \gg \mathbf{0}$ *such that* $\mathbf{Ax}^0 = \lambda_0 \mathbf{x}^0$; (b) *if* $\lambda \neq \lambda_0$ *is any other eigenvalue of* $\mathbf{A}$, *then* $|\lambda| < \lambda_0$; (c) *the right eigenvectors of* $\mathbf{A}$ *with eigenvalue* λ_0 *form a one-dimensional subspace, i.e.,* $\dim \mathfrak{W}_{\lambda_0} = 1$.

Proof. (a) By definition of λ_0, there exists a sequence $\gamma_1, \gamma_2, \ldots \to \lambda_0$ and vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ such that

$$\mathbf{x}^{(i)} > \mathbf{0}, \qquad \mathbf{A}\mathbf{x}^{(i)} \geq \gamma_i \mathbf{x}^{(i)}, \qquad \text{and} \qquad x_1^{(i)} + \cdots + x_n^{(i)} = 1. \quad \text{(A2.1)}$$

Since the components of all the $\mathbf{x}^{(i)}$ lie in the interval [0, 1], we can determine by a diagonalization procedure a sequence of positive integers $n_1 < n_2 < n_3 < \cdots$ and a vector $\mathbf{x}^0 = (x_1^0, \ldots, x_n^0)$ where $x_r^0 \in [0, 1]$ $(r = 1, 2, \ldots)$ such that

$$\lim_{j \to \infty} x_r^{(n_j)} \to x_r^0, \qquad r = 1, \ldots, n. \quad \text{(A2.2)}$$

It follows from (2.1) that $x_1^0 + \cdots + x_n^0 = 1$ and that $\mathbf{x}^0 > \mathbf{0}$. Furthermore, if we replace i by n_j in the second inequality of (2.1) and let $j \to \infty$, it follows that $\mathbf{A}\mathbf{x}^0 \geq \lambda_0 \mathbf{x}^0$. We claim that in fact $\mathbf{A}\mathbf{x}^0 = \lambda_0 \mathbf{x}$, for otherwise $\mathbf{A}\mathbf{x}^0 > \lambda_0 \mathbf{x}^0$. But then, applying $\mathbf{A}$ to both sides of this last inequality, remembering that $\mathbf{A} \gg \mathbf{0}$, and setting $\mathbf{y}^0 = \mathbf{A}\mathbf{x}^0$, we infer that $\mathbf{A}\mathbf{y}^0 \gg \lambda_0 \mathbf{y}^0$ and $\mathbf{y}^0 \gg \mathbf{0}$. Thus for $\varepsilon > 0$ and sufficiently small we have $\mathbf{A}\mathbf{y}^0 \gg (\lambda_0 + \varepsilon)\mathbf{y}^0$; multiplying $\mathbf{y}^0$ by a suitable positive constant so as to make the sum of its components equal to 1, we see that $\lambda_0 + \varepsilon$ belongs to Λ, which contradicts the definition of λ_0. Hence $\mathbf{A}\mathbf{x}^0 = \lambda_0 \mathbf{x}^0$. Since $\mathbf{x}^0 > \mathbf{0}$ and $\mathbf{A} \gg \mathbf{0}$, we have $\lambda_0 \mathbf{x}^0 \gg \mathbf{0}$, or $\mathbf{x}^0 \gg \mathbf{0}$, which proves (a).

(b) Suppose that $\lambda \neq \lambda_0$ and $\mathbf{A}\mathbf{z} = \lambda \mathbf{z}$, where $\mathbf{z} \neq \mathbf{0}$. In component form $\mathbf{A}\mathbf{z} = \lambda \mathbf{z}$ is just

$$\sum_{j=1}^{n} a_{ij} z_j = \lambda z_i, \qquad i = 1, \ldots, n.$$

Taking the absolute value of both sides, keeping in mind that $a_{ij} > 0$, and using the fact that the absolute value of a sum does not exceed the sum of the absolute values of the summands, we obtain

$$\sum_{j=1}^{n} a_{ij} |z_j| \geq |\lambda| \, |z_i|, \qquad i = 1, \ldots, n,$$

i.e.,

$$\mathbf{A}|\mathbf{z}| \geq |\lambda| \, |\mathbf{z}|, \qquad \text{where} \quad |\mathbf{z}| = (|z_1|, \ldots, |z_n|).$$

Multiplying $|\mathbf{z}|$ by a suitable positive constant so that the sum of the components is equal to 1 (recall that $\mathbf{z} \neq 0$), we see that $|\lambda|$ belongs to Λ. Thus $|\lambda| \leq \lambda_0$ by the definition of λ_0. To show that $|\lambda| < \lambda_0$, consider the matrix $\mathbf{A}_\delta = \mathbf{A} - \delta \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and δ is chosen so small that $\mathbf{A}_\delta \gg \mathbf{0}$. Since λ_0 is the largest positive eigenvalue of $\mathbf{A}$, $\lambda_0 - \delta$ is the largest positive eigenvalue of $\mathbf{A}_\delta$.

We repeat the above argument for $|\lambda| \leq \lambda_0$ with $\mathbf{A}$ and λ replaced by

$\mathbf{A}_\delta$ and $\lambda - \delta$, respectively. It follows that $|\lambda - \delta| \leq \lambda_0 - \delta$. But

$$|\lambda| = |\lambda - \delta + \delta| \leq |\lambda - \delta| + \delta \leq \lambda_0,$$

so that $|\lambda| = \lambda_0$ implies $|\lambda| = |\lambda - \delta| + \delta$, which requires that λ be real and positive. Therefore $\lambda = |\lambda| = \lambda_0$, and this contradicts the assumption $\lambda \neq \lambda_0$.

(c) Suppose that $\mathbf{Ay} = \lambda_0 \mathbf{y}$ and for no constant c does $\mathbf{y} = c\mathbf{x}^0$. Since $\mathbf{A}$ is a real matrix the vectors $\mathbf{u}$, $\mathbf{v}$ whose components consist of the real and imaginary parts, respectively, of the components of $\mathbf{y}$ are also eigenvectors of $\mathbf{A}$ with eigenvalue λ_0, and since $\mathbf{y} \neq c\mathbf{x}^0$ for any value of c, it follows that at least one of $\mathbf{u}$, $\mathbf{v}$ is not of the form $c\mathbf{x}^0$. Thus, we might just as well assume $\mathbf{y}$ to be real in the first place. As $\mathbf{x}^0 \gg 0$, we can choose μ so that $\mathbf{x}^0 - \mu\mathbf{y} > \mathbf{0}$ but *not* $\gg \mathbf{0}$; that is to say, we can take $|\mu| = \min_{y_i \neq 0}\{x_i^0/|y_i|\}$, and of appropriate sign. But $\mathbf{A}(\mathbf{x}^0 - \mu\mathbf{y}) = \lambda_0(\mathbf{x}^0 - \mu\mathbf{y})$; as in the proof of part (a), necessarily $(\mathbf{x}^0 - \mu\mathbf{y}) \gg 0$, which is in contradiction to the choice of μ. ∎

Before continuing, let us make a few simple observations. If $\mathbf{A} \gg \mathbf{0}$, then we can assert the existence of a vector $\mathbf{f}^0 \gg \mathbf{0}$ such that $\mathbf{f}^0\mathbf{A} = \lambda_0\mathbf{f}^0$, and the manifold of left eigenvectors corresponding to λ_0 is one-dimensional. To verify this let $\lambda' = \sup_\Lambda \lambda$ where $\Lambda' = \{\lambda | \mathbf{fA} \geq \lambda\mathbf{f}$ for some $\mathbf{f} > \mathbf{0}\}$, and as in the proof of Theorem 2.1, we conclude the existence of $\mathbf{f}^0 \gg \mathbf{0}$ such that $\mathbf{f}^0\mathbf{A} = \lambda'\mathbf{f}^0$, $|\lambda| < \lambda'$ if λ is any eigenvalue $\neq \lambda'$ and the manifold of the left eigenvectors corresponding to λ' is one-dimensional. But this implies that $|\lambda_0| < \lambda'$ if $\lambda_0 \neq \lambda'$, since λ_0 is an eigenvalue of $\mathbf{A}$. But Theorem 2.1 says that $|\lambda'| < \lambda_0$ if the eigenvalue λ' is different from λ_0. Hence $\lambda' = \lambda_0$.

Theorem 2.2. *If $\mathbf{A} > \mathbf{0}$ and $\mathbf{A}^m \gg \mathbf{0}$ for some integer $m > 0$, then the assertions of the preceding theorem hold.*

Proof. As in the proof of Theorem 2.1, we can find $\mathbf{x}^0 > \mathbf{0}$ such that $\mathbf{Ax}^0 \geq \lambda_0\mathbf{x}^0$. If $\mathbf{Ax}^0 \neq \lambda_0\mathbf{x}^0$, then $\mathbf{Ax}^0 > \lambda_0\mathbf{x}^0$. Applying $\mathbf{A}^m$ to both sides, we find that $\mathbf{A}^{m+1}\mathbf{x}^0 \gg \lambda_0\mathbf{A}^m\mathbf{x}^0$, and $\mathbf{y} = \mathbf{A}^m\mathbf{x}^0 \gg 0$. Thus $\mathbf{Ay} \gg \lambda_0\mathbf{y}$ and, by the proof of Theorem 2.1, this contradicts the definition of λ_0; hence $\mathbf{Ax}^0 = \lambda_0\mathbf{x}^0$. Applying $\mathbf{A}$ successively to both sides of $\mathbf{Ax}^0 = \lambda_0\mathbf{x}^0$, we find $\mathbf{A}^m\mathbf{x}^0 = \lambda_0^m\mathbf{x}^0$. Since $\mathbf{A}^m \gg \mathbf{0}$ and $\mathbf{x}^0 > \mathbf{0}$, we conclude that $\lambda_0^m\mathbf{x}^0 \gg \mathbf{0}$, and hence $\mathbf{x}^0 \gg \mathbf{0}$. For Theorem 2.1(b), the proof that $|\lambda| \leq \lambda_0$ depended only upon $\mathbf{A} > \mathbf{0}$. Suppose then, that $|\lambda| = \lambda_0$, and $\mathbf{Az} = \lambda\mathbf{z}$ for some $\mathbf{z} \neq \mathbf{0}$. Then $\mathbf{A}^m\mathbf{z} = \lambda^m\mathbf{z}$, $\mathbf{A}^m\mathbf{x}^0 = \lambda_0^m\mathbf{x}^0$, and $|\lambda^m| = \lambda_0^m$. If we knew that λ_0^m was the largest positive eigenvalue of $\mathbf{A}^m$, the proof would paraphrase that of Theorem 2.1. Now, since $\mathbf{A}^m \gg \mathbf{0}$ by Theorem 2.1 we know

that $\mathbf{A}^m$ has a largest positive eigenvalue with a corresponding eigenvector all of whose components are positive. Thus, if λ_0^m is not the largest positive eigenvalue of $\mathbf{A}^m$, we may conclude that $\mathbf{A}^m$ has two positive eigenvalues $\lambda_1 > \lambda_2$ and corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2 \gg 0$. But this is not possible; for, let $\mu > 0$ be such that $\mathbf{x}_2 - \mu\mathbf{x}_1 > 0$ but not $\gg \mathbf{0}$. Then $\mathbf{A}^m(\mathbf{x}_2 - \mu\mathbf{x}_1) \gg 0$. On the other hand, $\mathbf{A}^m(x_2 - \mu\mathbf{x}_1) = \lambda_2\mathbf{x}_2 - \mu\lambda_1\mathbf{x}_1$ $= \lambda_2(\mathbf{x}_2 - \mu\mathbf{x}_1) - (\lambda_1 - \lambda_2)\mu\mathbf{x}_1$. Since the first term is *not* $\gg 0$ while the second term is $\gg 0$, we obtain a contradiction. The proof of (c) is identical with that given for Theorem 2.1(c), once one observes that any eigenvector of $\mathbf{A}$ is an eigenvector of $\mathbf{A}^m$. ∎

To continue our study of matrices $\mathbf{A} > \mathbf{0}$ for which $\mathbf{A}^m \gg \mathbf{0}$ for some integer $m > 0$, we introduce a matrix of rank 1 of the form

$$\mathbf{P} = \|x_i^0 f_j^0\|,$$

where $\mathbf{x}^0$ is the same as earlier, and $\mathbf{f}^0 \gg \mathbf{0}$ satisfies $\mathbf{f}^0\mathbf{A} = \lambda_0\mathbf{f}^0$ and is normalized by a multiplicative factor so that $\sum_{i=1}^n x_i^0 f_i^0 = 1$. Then $\mathbf{P}$ has the following properties:

(i) For any vectors $\mathbf{x}$ and $\mathbf{f}$, $\mathbf{Px} = (\mathbf{x}, \mathbf{f}^0)\mathbf{x}^0$, $\mathbf{fP} = (\mathbf{f}, \mathbf{x}^0)\mathbf{f}^0$, in particular $\mathbf{Px}^0 = \mathbf{x}^0$, $\mathbf{f}^0\mathbf{P} = \mathbf{f}^0$.

(ii) $\mathbf{P}^2 = \mathbf{P}$.

(iii) $\mathbf{AP} = \mathbf{PA} = \lambda_0\mathbf{P}$.

The first two assertions are verified by direct computation; as for the third, we observe that for any vector $\mathbf{x}$

$$\mathbf{APx} = \mathbf{A}(\mathbf{x}, \mathbf{f}^0)\mathbf{x}^0 = (\mathbf{x}, \mathbf{f}^0)\mathbf{Ax}^0 = (\mathbf{x}, \mathbf{f}^0)\lambda_0\mathbf{x}^0 = \lambda_0\mathbf{Px},$$

so that $\mathbf{AP} = \lambda_0\mathbf{P}$; similarly $\mathbf{fPA} = \mathbf{f}\lambda_0\mathbf{P}$, which implies $\mathbf{PA} = \lambda_0\mathbf{P}$.

We now quote without proof the following fact: Let $\mathbf{B}$ be a (square) matrix; set $\mathbf{B}^n = \|b_{ij}^{(n)}\|$, and

$$r = \max_{i,j} \overline{\lim_{n\to\infty}} \sqrt[n]{|b_{ij}^{(n)}|}.$$

Then $\mathbf{B}$ has an eigenvalue λ^* such that $|\lambda^*| = r$, and if λ is any other eigenvalue of $\mathbf{B}$, then $|\lambda| \leq r$. Frequently r is called the spectral radius of $\mathbf{B}$. We are now prepared to prove the following theorem.

Theorem 2.3. *If* $\mathbf{A} > \mathbf{0}$, $\mathbf{A}^m \gg \mathbf{0}$ *for some integer* $m > 0$, *and* λ_0 *and* $\mathbf{P}$ *are defined as above, then*

$$\frac{1}{\lambda_0^n}\mathbf{A}^n \to \mathbf{P} \qquad \text{as} \quad n \to \infty.$$

Proof. We assert first that if λ is an eigenvalue of $\mathbf{B} = \mathbf{A} - \lambda_0\mathbf{P}$, then $|\lambda| < \lambda_0$. In fact, suppose that $\mathbf{Bz} = \lambda\mathbf{z}$ for some $\mathbf{z} \neq \mathbf{0}$. Then

$$\lambda\mathbf{Pz} = \mathbf{PBz} = \mathbf{P}(\mathbf{A} - \lambda_0\mathbf{P})\mathbf{z} = (\lambda_0\mathbf{P} - \lambda_0\mathbf{P}^2)\mathbf{z} = \lambda_0(\mathbf{P} - \mathbf{P})\mathbf{z} = \mathbf{0},$$

and so $\mathbf{Bz} = \lambda\mathbf{z}$ reduces to $\mathbf{Az} = \lambda\mathbf{z}$. We know from Theorem 2.2 that either $\lambda = \lambda_0$ or $|\lambda| < \lambda_0$. If $\lambda = \lambda_0$, then $\mathbf{Az} = \lambda_0\mathbf{z}$, and therefore $\mathbf{z}$ is a multiple of $\mathbf{x}^0$. But as established above, $\lambda\mathbf{Pz} = \lambda\mathbf{z} \neq \mathbf{0}$ is impossible. Hence the spectral radius r of $\mathbf{B}$ satisfies $r < \lambda_0$. Let ρ satisfy $r < \rho < \lambda_0$; since

$$r = \overline{\lim_{n\to\infty}} \sqrt[n]{\max_{i,j}|b_{ij}^{(n)}|} < \rho,$$

it follows that $\max_{i,j}|b_{ij}^{(n)}| < \rho^n$ for n sufficiently large. Using properties (ii) and (iii) of $\mathbf{P}$, we readily verify by induction that

$$\mathbf{B}^m = \mathbf{A}^m - \lambda_0^m\mathbf{P}$$

or

$$\frac{\mathbf{A}^m}{\lambda_0^m} = \frac{\mathbf{B}^m}{\lambda_0^m} + \mathbf{P}.$$

Since $\max_{i,j}|b_{ij}^{(n)}| < \rho^n$ for n sufficiently large,

$$\left|\frac{b_{ij}^{(m)}}{\lambda_0^m}\right| \leq \left(\frac{\rho}{\lambda_0}\right)^m \to 0,$$

and so $\mathbf{B}^m/\lambda^m \to 0$. ∎

The Second Frobenius Theorem. The main Frobenius theorem is as follows.

Theorem 2.4. *Assume* $\mathbf{A} > \mathbf{0}$, *and let* λ_0 *be defined as in Theorem* 2.1. *Then* (a) λ_0 *is an eigenvalue of* $\mathbf{A}$ *with an eigenvector* $\mathbf{x}^0 > \mathbf{0}$; (b) *if* λ *is any other eigenvector of* $\mathbf{A}$, *then* $|\lambda| \leq \lambda_0$; (c)

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\mathbf{A}^i}{\lambda_0^i}$$

converges if $\mathbf{x}^0 \gg \mathbf{0}$; (d) *if* λ *is an eigenvalue of* $\mathbf{A}$ *and* $|\lambda| = \lambda_0$, *then* $\eta = \lambda/\lambda_0$ *is a root of unity and* $\eta^m\lambda_0$ *is an eigenvalue of* $\mathbf{A}$ *for* $m = 0, 1, 2, \ldots$.

Proof. (a) Let $\mathbf{E}$ be the matrix all of whose components equal 1; hence $\mathbf{A} + \delta\mathbf{E} \gg \mathbf{0}$ for every $\delta > 0$. Let $0 < \delta_1 < \delta_2$; and choose $\mathbf{x} = (x_1, \ldots, x_n) > 0$ such that $\sum_{i=1}^n x_i = 1$. Then $(\mathbf{A} + \delta_1\mathbf{E})\mathbf{x} \geq \lambda\mathbf{x}$ implies that

$$\begin{aligned}(\mathbf{A} + \delta_2\mathbf{E})\mathbf{x} &= (\mathbf{A} + \delta_1\mathbf{E})\mathbf{x} + (\delta_2 - \delta_1)\mathbf{E}\mathbf{x} \\ &\geq [\lambda + (\delta_2 - \delta_1)]\mathbf{x}.\end{aligned}$$

Thus, if $\lambda_0(\delta)$ is the value of λ_0 corresponding to the matrix $\mathbf{A}+\delta\mathbf{E}$, we see that $\lambda_0(\delta)$ is an increasing function of δ. We note that $\lambda_0(0)$ is the value of λ_0 corresponding to $\mathbf{A}$ itself. Now Theorem 2.1 affirms the existence of a vector $\mathbf{x}(\delta) \gg 0$, normalized so that $\sum_{i=1}^{n} x_i(\delta) = 1$ and satisfying

$$(\mathbf{A}+\delta\mathbf{E})\mathbf{x}(\delta) = \lambda_0(\delta)\mathbf{x}(\delta).$$

Let $\delta_1 > \delta_2 > \cdots$ be a positive sequence whose limit is zero. By the proof of Theorem 2.1, we can find integers $n_1, n_2, \ldots$ such that $\lim_{j\to\infty} \mathbf{x}(\delta_{n_j}) \to \mathbf{x}^0$, where $\mathbf{x}^0$ is some vector > 0 and $\sum_{i=1}^{n} x_i^0 = 1$. Clearly $\mathbf{A}+\delta_{n_j}\mathbf{E} \to \mathbf{A}$ and $\lambda_0(\delta_{n_j}) \to \lambda' \geq \lambda_0$. Since

$$(\mathbf{A}+\delta_{n_j}\mathbf{E})\mathbf{x}(\delta_{n_j}) = \lambda_0(\delta_{n_j})\mathbf{x}(\delta_{n_j}),$$

letting $j \to \infty$ yields $\mathbf{A}\mathbf{x}^0 = \lambda'\mathbf{x}^0$. But according to the characterization of λ_0 established in Theorem 2.1(b), $\lambda_0 \geq \lambda'$; hence $\lambda_0 = \lambda'$ and (a) is proved.

The proof of (b) is identical to that for the case $\mathbf{A} \gg \mathbf{0}$.

For (c) and (d), there is clearly no loss of generality in assuming $\lambda_0 = 1$, since otherwise we could divide every element of $\mathbf{A}$ by λ_0.

(c) Since $\mathbf{A}\mathbf{x}_0 = \mathbf{x}_0$ we have $\mathbf{A}^m\mathbf{x}_0 = \mathbf{x}_0$. Writing this equation in terms of components we find immediately that

$$0 \leq a_{ij}^{(m)} \leq \frac{\max_i x_i^0}{\min_i x_i^0}.$$

Hence the elements of $\mathbf{A}^m$ are uniformly bounded.

Let $\mathbf{L} = \{\mathbf{x} | \mathbf{A}\mathbf{x} = \mathbf{x}\}$ and $\mathbf{K} = \{\mathbf{y} | \mathbf{y} = (\mathbf{I}-\mathbf{A})\mathbf{x} \text{ for some } \mathbf{x}\}$; i.e., $\mathbf{L}$ is the linear space of fixed points of $\mathbf{A}$, and $\mathbf{K}$ is the linear space of the range of the matrix $\mathbf{I}$–$\mathbf{A}$. In addition define

$$\mathbf{S}_m = \frac{\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^m}{m}.$$

Clearly $\mathbf{L}$ is a closed linear space such that, for every $\mathbf{x}$ in $\mathbf{L}$,

$$\mathbf{S}_m\mathbf{x} = \frac{\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^m}{m}\mathbf{x} = \mathbf{x},$$

and so $\lim_{m\to\infty} \mathbf{S}_m\mathbf{x} = \mathbf{x}$. We will show that $\mathbf{S}_m\mathbf{x}$ also converges for every $\mathbf{x}$ in $\mathbf{K}$ and that every vector $\mathbf{x}$ in n-dimensional coordinate space is in $\mathbf{L} \oplus \mathbf{K}$, (the direct sum of the spaces $\mathbf{L}$ and $\mathbf{K}$). This will complete the proof of (c).

We assert that for $\mathbf{y} \in \mathbf{K}$, $\lim_{m\to\infty} \mathbf{S}_m\mathbf{y} = \mathbf{0}$. In fact, since $\mathbf{y} = (\mathbf{I} - \mathbf{A})\mathbf{x}$ for some $\mathbf{x}$,

$$\mathbf{S}_m\mathbf{y} \equiv \frac{\mathbf{A}\mathbf{y} + \cdots + \mathbf{A}^m\mathbf{y}}{m} = \frac{\mathbf{A}\mathbf{x} - \mathbf{A}^{m+1}\mathbf{x}}{m}$$

tends to $\mathbf{0}$ as $m \to \infty$ by virtue of the fact that the elements of $\mathbf{A}^m$ are uniformly bounded.

To show that every vector $\mathbf{x}$ is the sum of a vector in $\mathbf{L}$ and a vector in $\mathbf{K}$ consider

$$\mathbf{x} = (\mathbf{x} - \mathbf{S}_m\mathbf{x}) + \mathbf{S}_m\mathbf{x} = \mathbf{y}_m + \mathbf{z}_m .$$

Since the elements of $\mathbf{A}^m$ are uniformly bounded we may conclude that the components of $\mathbf{y}_m$ and $\mathbf{z}_m$ are also bounded. Hence there exists a sequence of positive integers $m_1 < m_2 < \cdots$ and a vector $\mathbf{z}^0$ such that

$$\lim_{i\to\infty} \mathbf{z}_{m_i} = \mathbf{z}_0$$

Since

$$\mathbf{z}_{m_i} - \mathbf{A}\mathbf{z}_{m_i} = \frac{\mathbf{A} - \mathbf{A}^{m_i+1}}{m_i}\mathbf{x} \to \mathbf{0} \qquad \text{as} \quad i \to \infty,$$

we have

$$\mathbf{z}_0 = \lim_{i\to\infty} \mathbf{z}_{m_i} = \lim_{i\to\infty} \mathbf{A}\mathbf{z}_{m_i} = \mathbf{A} \lim_{i\to\infty} \mathbf{z}_{m_i} = \mathbf{A}\mathbf{z}_0$$

and $\mathbf{z}_0 \in \mathbf{L}$.

Also

$$\mathbf{y}_m = \mathbf{x} - \mathbf{S}_m\mathbf{x} = \frac{1}{m}\left[(\mathbf{x} - \mathbf{A}\mathbf{x}) + (\mathbf{x} - \mathbf{A}^2\mathbf{x}) + \cdots + (\mathbf{x} - \mathbf{A}^m\mathbf{x})\right]$$

$$= (\mathbf{I} - \mathbf{A})\left[\frac{\mathbf{x}}{m} + \frac{(\mathbf{I} + \mathbf{A})\mathbf{x}}{m} + \frac{(\mathbf{I} + \mathbf{A} + \mathbf{A}^2)\mathbf{x}}{m} + \cdots + \right.$$

$$\left. + \frac{(\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{m-1})\mathbf{x}}{m}\right],$$

which implies that $\mathbf{y}_m \in \mathbf{K}$. Since $\mathbf{K}$ is a closed linear space and the elements of $\mathbf{y}_m$ are uniformly bounded, $\mathbf{y}_{m_i} \to \mathbf{x} - \mathbf{z}_0 \in \mathbf{K}$ as $i \to \infty$. Thus $\mathbf{x} = (\mathbf{x} - \mathbf{z}_0) + \mathbf{z}_0$ where $\mathbf{x} - \mathbf{z}_0 \in \mathbf{K}$, $\mathbf{z}_0 \in \mathbf{L}$, and the proof is complete.

(d) We know that there exists a vector $\mathbf{f}^0 > 0$ such that $\mathbf{f}^0\mathbf{A} = \mathbf{f}^0$. Let us assume first that $\mathbf{f}^0 \gg \mathbf{0}$. Suppose now that $\lambda \neq 1$, $|\lambda| = 1$, and that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$. Then

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i \qquad i = 1, 2, \ldots, n$$

and so

$$\sum_{j=1}^{n} a_{ij}|x_j| \geq |x_i| \qquad \text{or} \qquad \mathbf{A}|\mathbf{x}| \geq |\mathbf{x}|.$$

But if $\mathbf{A}|\mathbf{x}| > |\mathbf{x}|$ then $(\mathbf{f}^0, |\mathbf{x}|) < (\mathbf{f}^0, \mathbf{A}|\mathbf{x}|) = (\mathbf{f}^0\mathbf{A}, |\mathbf{x}|) = (\mathbf{f}^0, |\mathbf{x}|)$. This absurdity implies the result $\mathbf{A}|\mathbf{x}| = |\mathbf{x}|$. Consequently

$$\sum_{j=1}^{n} a_{ij}|x_j| = |x_i| = \left|\sum_{j=1}^{n} a_{ij}x_j\right|, \qquad i = 1, 2, \ldots, n$$

and so there exist constants $\mu_1, \ldots, \mu_n$, $|\mu_i| = 1$, such that

$$(*) \qquad a_{ij}x_j = a_{ij}|x_j|\mu_i \qquad \text{for all } i, j.$$

Let $\mathbf{x}\cdot\mathbf{y}$ denote the vector $(x_1y_1, \ldots, x_ny_n)$. Multiplying the preceding relation by μ_j^r (the rth power of μ_j) and summing over j, we find that

$$\mathbf{A}(\mathbf{x}\cdot\boldsymbol{\mu}^r) = \boldsymbol{\mu}\cdot\mathbf{A}(|\mathbf{x}|\cdot\boldsymbol{\mu}^r).$$

At the same time, summing over i in (*), we obtain

$$\mathbf{A}\mathbf{x} = \boldsymbol{\mu}\cdot\mathbf{A}|\mathbf{x}|,$$

from which follows

$$\lambda\mathbf{x} = \boldsymbol{\mu}\cdot|\mathbf{x}|.$$

Thus

$$\mathbf{A}(\mathbf{x}\cdot\boldsymbol{\mu}^r) = \boldsymbol{\mu}\cdot\mathbf{A}(\lambda\mathbf{x}\cdot\boldsymbol{\mu}^{r-1}) = \lambda\boldsymbol{\mu}\cdot\mathbf{A}(\mathbf{x}\cdot\boldsymbol{\mu}^{r-1}), \qquad r = 1, 2, \ldots,$$

from which it follows inductively that $\mathbf{A}(\mathbf{x}\cdot\boldsymbol{\mu}^r) = \lambda^{r+1}(\boldsymbol{\mu}^r\cdot\mathbf{x})$. Thus λ^r is an eigenvalue of $\mathbf{A}$ for $r = 1, 2, \ldots$. Since the number of eigenvalues of $\mathbf{A}$ is finite, λ must be a root of unity.

Suppose now that $\mathbf{f}^0 > \mathbf{0}$ but not $\gg \mathbf{0}$. By relabeling, if necessary, the rows and columns of $\mathbf{A}$, we may assume that $\mathbf{f}^0 = (f_1^0, \ldots, f_r^0, 0, \ldots, 0)$, where $f_i^0 > 0$, $i = 1, \ldots, r$. Since $\mathbf{A} > \mathbf{0}$ the relation $\mathbf{f}^0\mathbf{A} = \mathbf{f}^0$ implies the decomposition

$$\mathbf{A} = \begin{Vmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{B} & \mathbf{A}_2 \end{Vmatrix},$$

where $\mathbf{A}_1$ is an $r \times r$ matrix and $\mathbf{A}_2$ is an $n - r \times n - r$ matrix and also that the vector $(f_1^0, \ldots, f_r^0)$ is a left eigenvector, with eigenvalue 1, of $\mathbf{A}_1$. Let λ be an eigenvalue of $\mathbf{A}$; if λ is an eigenvalue of $\mathbf{A}_1$ we are back to the case considered with $\mathbf{A}_1$ replacing $\mathbf{A}$. If λ is not an eigenvalue of $\mathbf{A}_1$, it must be an eigenvalue of $\mathbf{A}_2$. But the eigenvalues of $\mathbf{A}_2$ are eigenvalues of $\mathbf{A}$ and they do not exceed 1 in absolute value. At the same time, since $\mathbf{A}_2 > \mathbf{0}$ it has a largest positive eigenvalue which is an upper bound for the absolute values of all its other eigenvalues. Since $|\lambda| = 1$, the largest positive eigenvalue of $\mathbf{A}_2$ is precisely 1. We can obviously apply to $\mathbf{A}_2$

the preceding analysis; either it has a left eigenvector $\gg \mathbf{0}$ corresponding to the eigenvalue 1, or else it is of the form (under suitable rearrangement of its rows and columns)

$$\mathbf{A}_2 = \begin{Vmatrix} \mathbf{A}_3 & \mathbf{0} \\ \mathbf{B}_1 & \mathbf{A}_4 \end{Vmatrix}.$$

Continuing in this way we can reduce the problem in a finite number of steps to the situation in which there exists a left eigenvector $\gg \mathbf{0}$ with eigenvalue 1. ∎

The following corollaries yield some useful information concerning the spectral radius $\lambda_0(\mathbf{A})$ of a positive matrix $\mathbf{A}$. The first corollary is simply a restatement of Theorem 2.4, (a) and (b).

Corollary 2.1. *If $\mathbf{A} > \mathbf{0}$, then the eigenvalue of largest magnitude $\lambda_0 = \lambda_0(\mathbf{A})$ is real and nonnegative and is characterized as $\lambda_0 = \max_\Lambda \lambda$, where*

$$\Lambda = \{\lambda | \mathbf{Ax} \geq \lambda \mathbf{x} \quad \text{for some} \quad \mathbf{x} > \mathbf{0}\}.$$

Corollary 2.2. *If $\mathbf{A} > \mathbf{0}$ and there exists $\mathbf{x}^0 \gg \mathbf{0}$ such that $\mathbf{Ax}^0 \leq \mu \mathbf{x}^0$, then μ is an upper bound for $\lambda_0(\mathbf{A})$.*

Proof. Applying $\mathbf{A}$ to both sides of $\mathbf{Ax}^0 \leq \mu \mathbf{x}^0$, we have $\mathbf{A}^2\mathbf{x}^0 \leq \mu \mathbf{Ax}^0 \leq \mu^2 \mathbf{x}^0$ and iterating we obtain

$$\mathbf{A}^n \mathbf{x}^0 \leq \mu^n \mathbf{x}^0, \qquad n = 1, 2, \ldots.$$

This readily implies

$$a_{ij}^{(n)} \leq \mu^n \frac{\max_i x_i^0}{\min_i x_i^0},$$

and so

$$\lambda_0(\mathbf{A}) = \overline{\lim_{n \to \infty}} \sqrt[n]{\max_{i,j} |a_{ij}^{(n)}|} \leq \mu. \quad \blacksquare$$

Corollary 2.3. *If $\mathbf{A} \geq \mathbf{B} \geq \mathbf{0}$, then $\lambda_0(\mathbf{B}) \leq \lambda_0(\mathbf{A})$.*

Proof. This can be seen either from Corollary 3.1 or from the relation

$$\lambda_0(\mathbf{A}) = \overline{\lim_{n \to \infty}} \sqrt[n]{\max_{i,j} a_{ij}^{(n)}} \geq \lim_{n \to \infty} \sqrt[n]{\max_{i,j} b_{ij}^{(n)}} = \lambda_0(\mathbf{B}),$$

since it is clear that $\mathbf{A} \geq \mathbf{B} \geq \mathbf{0}$ implies $\mathbf{A}^n \geq \mathbf{B}^n \geq \mathbf{0}$, $n = 1, 2, \ldots$. ∎

INDEX