7

Convergence of random variables

Summary. The many modes of convergence of a sequence of random variables are discussed and placed in context, and criteria are developed for proving convergence. These include standard inequalities, Skorokhod's theorem, the Borel–Cantelli lemmas, and the zero–one law. Laws of large numbers, including the strong law, are proved using elementary arguments. Martingales are defined, and the martingale convergence theorem proved, with applications. The relationship between prediction and conditional expectation is explored, and the condition of uniform integrability described.

7.1 Introduction

Expressions such as 'in the long run' and 'on the average' are commonplace in everyday usage, and express our faith that the averages of the results of repeated experimentation show less and less random fluctuation as they settle down to some limit.

(1) **Example. Buffon's needle (4.5.8).** In order to estimate the numerical value of π, Buffon devised the following experiment. Fling a needle a large number n of times onto a ruled plane and count the number S_n of times that the needle intersects a line. In accordance with the result of Example (4.5.8), the proportion S_n/n of intersections is found to be near to the probability $2/\pi$. Thus $X_n = 2n/S_n$ is a plausible estimate for π; this estimate converges as $n \to \infty$, and it seems reasonable to write $X_n \to \pi$ as $n \to \infty$. ●

(2) **Example. Decimal expansion.** Any number y satisfying $0 \leq y < 1$ has a decimal expansion

$$y = 0 \cdot y_1 y_2 \cdots = \sum_{j=1}^{\infty} y_j 10^{-j},$$

where each y_j takes some value in the set $\{0, 1, 2, \ldots, 9\}$. Now think of y_j as the outcome of a random variable Y_j where $\{Y_j\}$ is a family of independent variables each of which may take any value in $\{0, 1, 2, \ldots, 9\}$ with equal probability $\frac{1}{10}$. The quantity

$$Y = \sum_{j=1}^{\infty} Y_j 10^{-j}$$

is a random variable taking values in $[0, 1]$. It seems likely that Y is uniformly distributed on $[0, 1]$, and this turns out to be the case (see Problem (7.11.4)). More rigorously, this amounts to asserting that the sequence $\{X_n\}$ given by

$$X_n = \sum_{j=1}^{n} Y_j 10^{-j}$$

converges in some sense as $n \to \infty$ to a limit Y, and that this limit random variable is uniformly distributed on $[0, 1]$. ●

In both these examples we encountered a sequence $\{X_n\}$ of random variables together with the assertion that

$$X_n \to X \quad \text{as} \quad n \to \infty \tag{3}$$

for some other random variable X. However, random variables are real-valued functions on some sample space, and so (3) is a statement about the convergence of a sequence of *functions*. It is not immediately clear how such convergence is related to our experience of the theory of convergence of sequences $\{x_n\}$ of real numbers, and so we digress briefly to discuss sequences of functions.

Suppose for example that $f_1(\cdot), f_2(\cdot), \ldots$ is a sequence of functions mapping $[0, 1]$ into $\mathbb{R}$. In what manner may they converge to some limit function f?

(4) Convergence pointwise. If, for all $x \in [0, 1]$, the sequence $\{f_n(x)\}$ of real numbers satisfies $f_n(x) \to f(x)$ as $n \to \infty$ then we say that $f_n \to f$ *pointwise*. ●

(5) Norm convergence. Let V be a collection of functions mapping $[0, 1]$ into $\mathbb{R}$, and assume V is endowed with a function $\|\cdot\| : V \to \mathbb{R}$ satisfying:

(a) $\|f\| \geq 0$ for all $f \in V$,

(b) $\|f\| = 0$ if and only if f is the zero function (or equivalent to it, in some sense to be specified),

(c) $\|af\| = |a| \cdot \|f\|$ for all $a \in \mathbb{R}$, $f \in V$,

(d) $\|f + g\| \leq \|f\| + \|g\|$ (this is called the *triangle inequality*).

The function $\|\cdot\|$ is called a *norm*. If $\{f_n\}$ is a sequence of members of V then we say that $f_n \to f$ *with respect to the norm* $\|\cdot\|$ if

$$\|f_n - f\| \to 0 \quad \text{as} \quad n \to \infty.$$

Certain special and important norms are given by the L_p norm

$$\|g\|_p = \left(\int_0^1 |g(x)|^p \, dx \right)^{1/p}$$

for $p \geq 1$ and any function g satisfying $\|g\|_p < \infty$. ●

(6) Convergence in measure. Let $\epsilon > 0$ be prescribed, and define the 'distance' between two functions $g, h : [0, 1] \to \mathbb{R}$ by

$$d_\epsilon(g, h) = \int_E dx$$

where $E = \{u \in [0, 1] : |g(u) - h(u)| > \epsilon\}$. We say that $f_n \to f$ *in measure* if

$$d_\epsilon(f_n, f) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for all } \epsilon > 0. \qquad \bullet$$

The convergence of $\{f_n\}$ according to one definition does not necessarily imply its convergence according to another. For example, we shall see later that:

(a) if $f_n \to f$ pointwise then $f_n \to f$ in measure, but the converse is not generally true,

(b) there exist sequences which converge pointwise but not with respect to $\|\cdot\|_1$, and vice versa.

In this chapter we shall see how to adapt these modes of convergence to suit families of *random variables*. Major applications of the ensuing theory include the study of the sequence

$$S_n = X_1 + X_2 + \cdots + X_n \tag{7}$$

of partial sums of an independent identically distributed sequence $\{X_i\}$; the law of large numbers of Section 5.10 will appear as a special case.

It will be clear, from our discussion and the reader's experience, that probability theory is indispensable in descriptions of many processes which occur naturally in the world. Often in such cases we are interested in the future values of the process, and thus in the long-term behaviour within the mathematical model; this is why we need to prove limit theorems for sequences of random variables. Many of these sequences are generated by less tractable operations than, say, the partial sums in (7), and general results such as the law of large numbers may not be enough. It turns out that many other types of sequence are guaranteed to converge; in particular we shall consider later the remarkable theory of 'martingales' which has important applications throughout theoretical and applied probability. This chapter continues in Sections 7.7 and 7.8 with a simple account of the convergence theorem for martingales, together with some examples of its use; these include the asymptotic behaviour of the branching process and provide rigorous derivations of certain earlier remarks (such as (5.4.6)). Conditional expectation is put on a firm footing in Section 7.9.

All readers should follow the chapter up to and including Section 7.4. The subsequent material may be omitted at the first reading.

Exercises for Section 7.1

1. Let $r \geq 1$, and define $\|X\|_r = \{\mathbb{E}|X^r|\}^{1/r}$. Show that:

(a) $\|cX\|_r = |c| \cdot \|X\|_r$ for $c \in \mathbb{R}$,

(b) $\|X + Y\|_r \leq \|X\|_r + \|Y\|_r$,

(c) $\|X\|_r = 0$ if and only if $\mathbb{P}(X = 0) = 1$.

This amounts to saying that $\|\cdot\|_r$ is a norm on the set of equivalence classes of random variables on a given probability space with finite rth moment, the equivalence relation being given by $X \sim Y$ if and only if $\mathbb{P}(X = Y) = 1$.

2. Define $\langle X, Y\rangle = \mathbb{E}(XY)$ for random variables X and Y having finite variance, and define $\|X\| = \sqrt{\langle X, X\rangle}$. Show that:

(a) $\langle aX + bY, Z\rangle = a\langle X, Z\rangle + b\langle Y, Z\rangle$,

(b) $\|X + Y\|^2 + \|X - Y\|^2 = 2(\|X\|^2 + \|Y\|^2)$, the *parallelogram property*,

(c) if $\langle X_i, X_j\rangle = 0$ for all $i \neq j$ then

$$\left\|\sum_{i=1}^{n} X_i\right\|^2 = \sum_{i=1}^{n} \|X_i\|^2.$$

3. Let $\epsilon > 0$. Let $g, h : [0,1] \to \mathbb{R}$, and define $d_\epsilon(g,h) = \int_E dx$ where $E = \{u \in [0,1] : |g(u) - h(u)| > \epsilon\}$. Show that d_ϵ does not satisfy the triangle inequality.

4. Lévy metric. For two distribution functions F and G, let

$$d(F,G) = \inf\{\delta > 0 : F(x-\delta) - \delta \le G(x) \le F(x+\delta) + \delta \text{ for all } x \in \mathbb{R}\}.$$

Show that d is a metric on the space of distribution functions.

5. Find random variables $X, X_1, X_2, \ldots$ such that $\mathbb{E}(|X_n - X|^2) \to 0$ as $n \to \infty$, but $\mathbb{E}|X_n| = \infty$ for all n.

7.2 Modes of convergence

There are four principal ways of interpreting the statement '$X_n \to X$ as $n \to \infty$'. Three of these are related to (7.1.4), (7.1.5), and (7.1.6), and the fourth is already familiar to us.

(1) Definition. Let $X, X_1, X_2, \ldots$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say:

(a) $X_n \to X$ **almost surely**, written $X_n \xrightarrow{\text{a.s.}} X$, if $\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$ is an event whose probability is 1,

(b) $X_n \to X$ **in rth mean**, where $r \ge 1$, written $X_n \xrightarrow{r} X$, if $\mathbb{E}|X_n^r| < \infty$ for all n and

$$\mathbb{E}(|X_n - X|^r) \to 0 \quad \text{as} \quad n \to \infty,$$

(c) $X_n \to X$ **in probability**, written $X_n \xrightarrow{\text{P}} X$, if

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for all } \epsilon > 0,$$

(d) $X_n \to X$ **in distribution**, written† $X_n \xrightarrow{\text{D}} X$, if

$$\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x) \quad \text{as} \quad n \to \infty$$

for all points x at which the function $F_X(x) = \mathbb{P}(X \le x)$ is continuous.

It is appropriate to make some remarks about the four sections of this potentially bewildering definition.

(a) The natural adaptation of Definition (7.1.4) is to say that $X_n \to X$ *pointwise* if the set $A = \{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$ satisfies $A = \Omega$. Such a condition is of little interest to probabilists since it contains no reference to probabilities. In part (a) of (1) we do not require that A is the whole of Ω, but rather that its complement A^c is a null set. There are several notations for this mode of convergence, and we shall use these later. They include

$$X_n \to X \textit{ almost everywhere}, \text{ or } X_n \xrightarrow{\text{a.e.}} X,$$
$$X_n \to X \textit{ with probability } 1, \text{ or } X_n \to X \text{ w.p.1}.$$

†Many authors avoid this notation since convergence in distribution pertains only to the *distribution function* of X and not to the variable X itself. We use it here for the sake of uniformity of notation, but refer the reader to note (d) below.

(b) It is easy to check by Minkowski's inequality (4.14.27) that

$$\|Y\|_r = (\mathbb{E}|Y^r|)^{1/r} = \left(\int |y|^r \, dF_Y \right)^{1/r}$$

defines a norm on the collection of random variables with finite rth moment, for any value of $r \geq 1$. Rewrite Definition (7.1.5) with this norm to obtain Definition (1b). Here we shall only consider positive integral values of r, though the subsequent theory can be extended without difficulty to deal with any real r not smaller than 1. Of most use are the values $r = 1$ and $r = 2$, in which cases we write respectively

$$X_n \xrightarrow{1} X, \text{ or } X_n \to X \textit{ in mean}, \text{ or l.i.m. } X_n = X,$$

$$X_n \xrightarrow{2} X, \text{ or } X_n \to X \textit{ in mean square}, \text{ or } X_n \xrightarrow{\text{m.s.}} X.$$

(c) The functions of Definition (7.1.6) had a common domain $[0, 1]$; the X_n have a common domain Ω, and the distance function d_ϵ is naturally adapted to become

$$d_\epsilon(Y, Z) = \mathbb{P}(|Y - Z| > \epsilon) = \int_E d\mathbb{P}$$

where $E = \{\omega \in \Omega : |Y(\omega) - Z(\omega)| > \epsilon\}$. This notation will be familiar to those readers with knowledge of the abstract integral of Section 5.6.

(d) We have seen this already in Section 5.9 where we discussed the continuity condition. Further examples of convergence in distribution are to be found in Chapter 6, where we saw, for example, that an irreducible ergodic Markov chain converges in distribution to its unique stationary distribution. Convergence in distribution is also termed *weak convergence* or *convergence in law*. Note that if $X_n \xrightarrow{D} X$ then $X_n \xrightarrow{D} X'$ for any X' which has the same distribution as X.

It is no surprise to learn that the four modes of convergence are not equivalent to each other. You may guess after some reflection that convergence in distribution is the weakest, since it is a condition only on the *distribution functions* of the X_n; it contains no reference to the *sample space* Ω and no information about, say, the dependence or independence of the X_n. The following example is a partial confirmation of this.

(2) Example. Let X be a Bernoulli variable taking values 0 and 1 with equal probability $\frac{1}{2}$. Let $X_1, X_2, \ldots$ be identical random variables given by $X_n = X$ for all n. The X_n are certainly not independent, but $X_n \xrightarrow{D} X$. Let $Y = 1 - X$. Clearly $X_n \xrightarrow{D} Y$ also, since X and Y have the same distribution. However, X_n cannot converge to Y in any other mode because $|X_n - Y| = 1$ always. ●

Cauchy convergence. As in the case of sequences of real numbers, it is often convenient to work with a definition of convergence which does not make explicit reference to the limit. For example, we say that the sequence $\{X_n : n \geq 1\}$ of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is *almost surely Cauchy convergent* if the set of points ω of the sample space for which the real sequence $\{X_n(\omega) : n \geq 1\}$ is Cauchy convergent is an event having probability 1, which is to say that

$$\mathbb{P}\Big(\{\omega \in \Omega : X_m(\omega) - X_n(\omega) \to 0 \text{ as } m, n \to \infty\}\Big) = 1.$$

(See Appendix I for a brief discussion of the Cauchy convergence of a sequence of real numbers.) Now, a sequence of reals converges if and only if it is Cauchy convergent. Thus, for any $\omega \in \Omega$, the real sequence $\{X_n(\omega) : n \geq 1\}$ converges if and only if it is Cauchy convergent, implying that $\{X_n : n \geq 1\}$ converges almost surely if and only if it is almost surely Cauchy convergent. Other modes of Cauchy convergence appear in Exercise (7.3.1) and Problem (7.11.11).

Here is the chart of implications between the modes of convergence. Learn it well. Statements such as

$$(X_n \xrightarrow{\text{P}} X) \Rightarrow (X_n \xrightarrow{\text{D}} X)$$

mean that any sequence which converges in probability also converges in distribution to the same limit.

(3) Theorem. *The following implications hold*:

$$\begin{matrix} (X_n \xrightarrow{\text{a.s.}} X) \\ (X_n \xrightarrow{r} X) \end{matrix} \nearrow\!\searrow \quad (X_n \xrightarrow{\text{P}} X) \Rightarrow (X_n \xrightarrow{\text{D}} X)$$

for any $r \geq 1$. *Also, if* $r > s \geq 1$ *then*

$$(X_n \xrightarrow{r} X) \Rightarrow (X_n \xrightarrow{s} X).$$

No other implications hold in general†.

The four basic implications of this theorem are of the general form 'if A holds, then B holds'. The converse implications are false in general, but become true if certain extra conditions are imposed; such partial converses take the form 'if B holds together with C, then A holds'. These two types of statement are sometimes said to be of the 'Abelian' and 'Tauberian' types, respectively; these titles are derived from the celebrated theory of the summability of series. Usually, there are many possible choices for appropriate sets C of extra conditions, and it is often difficult to establish attractive 'corrected converses'.

(4) Theorem.

(a) *If* $X_n \xrightarrow{\text{D}} c$, *where* c *is constant, then* $X_n \xrightarrow{\text{P}} c$.

(b) *If* $X_n \xrightarrow{\text{P}} X$ *and* $\mathbb{P}(|X_n| \leq k) = 1$ *for all* n *and some* k, *then* $X_n \xrightarrow{r} X$ *for all* $r \geq 1$.

(c) *If* $P_n(\epsilon) = \mathbb{P}(|X_n - X| > \epsilon)$ *satisfies* $\sum_n P_n(\epsilon) < \infty$ *for all* $\epsilon > 0$, *then* $X_n \xrightarrow{\text{a.s.}} X$.

You should become well acquainted with Theorems (3) and (4). The proofs follow as a series of lemmas. These lemmas contain some other relevant and useful results.

Consider briefly the first and principal part of Theorem (3). We may already anticipate some way of showing that convergence in probability implies convergence in distribution, since both modes involve probabilities of the form $\mathbb{P}(Y \leq y)$ for some random variable Y and real y. The other two implications require intermediate steps. Specifically, the relation between convergence in rth mean and convergence in probability requires a link between expectations and distributions. We have to move very carefully in this context; even apparently 'natural'

†But see (14).

statements may be false. For example, if $X_n \xrightarrow{\text{a.s.}} X$ (and therefore $X_n \xrightarrow{\text{P}} X$ also) then it does *not* necessarily follow that $\mathbb{E}X_n \to \mathbb{E}X$ (see (9) for an instance of this); this matter is explored fully in Section 7.10. The proof of the appropriate stage of Theorem (3) requires Markov's inequality (7).

(5) Lemma. *If* $X_n \xrightarrow{\text{P}} X$ *then* $X_n \xrightarrow{\text{D}} X$. *The converse assertion fails in general*†.

Proof. Suppose $X_n \xrightarrow{\text{P}} X$ and write

$$F_n(x) = \mathbb{P}(X_n \le x), \quad F(x) = \mathbb{P}(X \le x),$$

for the distribution functions of X_n and X respectively. If $\epsilon > 0$,

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \le x) = \mathbb{P}(X_n \le x,\ X \le x+\epsilon) + \mathbb{P}(X_n \le x,\ X > x+\epsilon) \\ &\le F(x+\epsilon) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Similarly,

$$\begin{aligned} F(x-\epsilon) &= \mathbb{P}(X \le x-\epsilon) = \mathbb{P}(X \le x-\epsilon,\ X_n \le x) + \mathbb{P}(X \le x-\epsilon,\ X_n > x) \\ &\le F_n(x) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Thus

$$F(x-\epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \le F_n(x) \le F(x+\epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Let $n \to \infty$ to obtain

$$F(x-\epsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x+\epsilon)$$

for all $\epsilon > 0$. If F is continuous at x then

$$F(x-\epsilon) \uparrow F(x) \quad \text{and} \quad F(x+\epsilon) \downarrow F(x) \quad \text{as} \quad \epsilon \downarrow 0,$$

and the result is proved. Example (2) shows that the converse is false. ■

(6) Lemma.

(a) *If* $r > s \ge 1$ *and* $X_n \xrightarrow{r} X$ *then* $X_n \xrightarrow{s} X$.

(b) *If* $X_n \xrightarrow{1} X$ *then* $X_n \xrightarrow{\text{P}} X$.

The converse assertions fail in general.

This includes the fact that convergence in rth mean implies convergence in probability. Here is a useful inequality which we shall use in the proof of this lemma.

(7) Lemma. Markov's inequality. *If* X *is any random variable with finite mean then*

$$\mathbb{P}(|X| \ge a) \le \frac{\mathbb{E}|X|}{a} \quad \textit{for any } a > 0.$$

†But see (14).

Proof. Let $A = \{|X| \geq a\}$. Then $|X| \geq aI_A$ where I_A is the indicator function of A. Take expectations to obtain the result. ■

Proof of Lemma (6).

(a) By the result of Problem (4.14.28),

$$\left[\mathbb{E}(|X_n - X|^s)\right]^{1/s} \leq \left[\mathbb{E}(|X_n - X|^r)\right]^{1/r}$$

and the result follows immediately. To see that the converse fails, define an independent sequence $\{X_n\}$ by

$$X_n = \begin{cases} n & \text{with probability } n^{-\frac{1}{2}(r+s)}, \\ 0 & \text{with probability } 1 - n^{-\frac{1}{2}(r+s)}. \end{cases} \tag{8}$$

It is an easy *exercise* to check that

$$\mathbb{E}|X_n^s| = n^{\frac{1}{2}(s-r)} \to 0, \quad \mathbb{E}|X_n^r| = n^{\frac{1}{2}(r-s)} \to \infty.$$

(b) By Markov's inequality (7),

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}|X_n - X|}{\epsilon} \quad \text{for all } \epsilon > 0$$

and the result follows immediately. To see that the converse fails, define an independent sequence $\{X_n\}$ by

$$X_n = \begin{cases} n^3 & \text{with probability } n^{-2}, \\ 0 & \text{with probability } 1 - n^{-2}. \end{cases} \tag{9}$$

Then $\mathbb{P}(|X| > \epsilon) = n^{-2}$ for all large n, and so $X_n \xrightarrow{\text{P}} 0$. However, $\mathbb{E}|X_n| = n \to \infty$. ■

(10) Lemma. *Let* $A_n(\epsilon) = \{|X_n - X| > \epsilon\}$ *and* $B_m(\epsilon) = \bigcup_{n \geq m} A_n(\epsilon)$. *Then*:

(a) $X_n \xrightarrow{\text{a.s.}} X$ *if and only if* $\mathbb{P}(B_m(\epsilon)) \to 0$ *as* $m \to \infty$, *for all* $\epsilon > 0$,

(b) $X_n \xrightarrow{\text{a.s.}} X$ *if* $\sum_n \mathbb{P}(A_n(\epsilon)) < \infty$ *for all* $\epsilon > 0$,

(c) *if* $X_n \xrightarrow{\text{a.s.}} X$ *then* $X_n \xrightarrow{\text{P}} X$, *but the converse fails in general.*

Proof.

(a) Let $C = \{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$ and let

$$A(\epsilon) = \{\omega \in \Omega : \omega \in A_n(\epsilon) \text{ for infinitely many values of } n\} = \bigcap_m \bigcup_{n=m}^{\infty} A_n(\epsilon).$$

Now $X_n(\omega) \to X(\omega)$ if and only if $\omega \notin A(\epsilon)$ for all $\epsilon > 0$. Hence $\mathbb{P}(C) = 1$ implies $\mathbb{P}(A(\epsilon)) = 0$ for all $\epsilon > 0$. On the other hand, if $\mathbb{P}(A(\epsilon)) = 0$ for all $\epsilon > 0$, then

$$\mathbb{P}(C^c) = \mathbb{P}\left(\bigcup_{\epsilon > 0} A(\epsilon)\right) = \mathbb{P}\left(\bigcup_{m=1}^{\infty} A(m^{-1})\right) \quad \text{since } A(\epsilon) \subseteq A(\epsilon') \text{ if } \epsilon \geq \epsilon'$$

$$\leq \sum_{m=1}^{\infty} \mathbb{P}(A(m^{-1})) = 0.$$

It follows that $\mathbb{P}(C) = 1$ if and only if $\mathbb{P}(A(\epsilon)) = 0$ for all $\epsilon > 0$.

In addition, $\{B_m(\epsilon) : m \geq 1\}$ is a decreasing sequence of events with limit $A(\epsilon)$ (see Problem (1.8.16)), and therefore $\mathbb{P}(A(\epsilon)) = 0$ if and only if $\mathbb{P}(B_m(\epsilon)) \to 0$ as $m \to \infty$.

(b) From the definition of $B_m(\epsilon)$,

$$\mathbb{P}(B_m(\epsilon)) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n(\epsilon))$$

and so $\mathbb{P}(B_m(\epsilon)) \to 0$ as $m \to \infty$ whenever $\sum_n \mathbb{P}(A_n(\epsilon)) < \infty$.

(c) We have that $A_n(\epsilon) \subseteq B_n(\epsilon)$, and therefore $\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(A_n(\epsilon)) \to 0$ whenever $\mathbb{P}(B_n(\epsilon)) \to 0$. To see that the converse fails, define an independent sequence $\{X_n\}$ by

$$\textbf{(11)} \qquad X_n = \begin{cases} 1 & \text{with probability } n^{-1}, \\ 0 & \text{with probability } 1 - n^{-1}. \end{cases}$$

Clearly $X_n \xrightarrow{\mathrm{P}} 0$. However, if $0 < \epsilon < 1$,

$$\begin{aligned} \mathbb{P}(B_m(\epsilon)) &= 1 - \lim_{r \to \infty} \mathbb{P}\big(X_n = 0 \text{ for all } n \text{ such that } m \leq n \leq r\big) && \text{by Lemma (1.3.5)} \\ &= 1 - \left(1 - \frac{1}{m}\right)\left(1 - \frac{1}{m+1}\right) \cdots && \text{by independence} \\ &= 1 - \lim_{M \to \infty} \left(\frac{m-1}{m} \frac{m}{m+1} \frac{m+1}{m+2} \cdots \frac{M}{M+1}\right) \\ &= 1 - \lim_{M \to \infty} \frac{m-1}{M+1} = 1 && \text{for all } m, \end{aligned}$$

and so $\{X_n\}$ does not converge almost surely. ■

(12) Lemma. *There exist sequences which*:

(a) *converge almost surely but not in mean,*

(b) *converge in mean but not almost surely.*

Proof.

(a) Consider Example (9). Use (10b) to show that $X_n \xrightarrow{\text{a.s.}} 0$.

(b) Consider Example (11). ■

This completes the proof of Theorem (3), and we move to Theorem (4).

Proof of Theorem (4).

(a) We have that

$$\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \to 0 \quad \text{if} \quad X_n \xrightarrow{\mathrm{D}} c.$$

(b) If $X_n \xrightarrow{\mathrm{P}} X$ and $\mathbb{P}(|X_n| \leq k) = 1$ then $\mathbb{P}(|X| \leq k) = 1$ also, since

$$\mathbb{P}(|X| \leq k + \epsilon) = \lim_{n \to \infty} \mathbb{P}(|X_n| \leq k + \epsilon) = 1$$

for all $\epsilon > 0$. Now, let $A_n(\epsilon) = \{|X_n - X| > \epsilon\}$, with complement $A_n(\epsilon)^c$. Then

$$|X_n - X|^r \leq \epsilon^r I_{A_n(\epsilon)^c} + (2k)^r I_{A_n(\epsilon)}$$

with probability 1. Take expectations to obtain

$$\mathbb{E}(|X_n - X|^r) \leq \epsilon^r + [(2k)^r - \epsilon^r]\mathbb{P}(A_n(\epsilon)) \to \epsilon^r \quad \text{as} \quad n \to \infty.$$

Let $\epsilon \downarrow 0$ to obtain that $X_n \xrightarrow{r} X$.

(c) This is just (10b). ■

Note that any sequence $\{X_n\}$ which satisfies $X_n \xrightarrow{P} X$ necessarily contains a subsequence $\{X_{n_i} : 1 \leq i < \infty\}$ which converges almost surely.

(13) Theorem. *If $X_n \xrightarrow{P} X$, there exists a non-random increasing sequence of integers $n_1, n_2, \ldots$ such that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$.*

Proof. Since $X_n \xrightarrow{P} X$, we have that

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty, \quad \text{for all } \epsilon > 0.$$

Pick an increasing sequence $n_1, n_2, \ldots$ of positive integers such that

$$\mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq i^{-2}.$$

For any $\epsilon > 0$,

$$\sum_{i > \epsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > \epsilon) \leq \sum_{i > \epsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > i^{-1}) < \infty$$

and the result follows from (10b). ■

We have seen that convergence in distribution is the weakest mode of convergence since it involves distribution functions only and makes no reference to an underlying probability space (see Theorem (5.9.4) for an equivalent formulation of convergence in distribution which involves distribution functions alone). However, assertions of the form '$X_n \xrightarrow{D} X$' (or equivalently '$F_n \to F$', where F_n and F are the distribution functions of X_n and X) have important and useful representations in terms of almost sure convergence.

(14) Skorokhod's representation theorem. *If $\{X_n\}$ and X, with distribution functions $\{F_n\}$ and F, are such that*

$$X_n \xrightarrow{D} X \text{ (or, equivalently, } F_n \to F) \text{ as } n \to \infty$$

then there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and random variables $\{Y_n\}$ and Y, mapping Ω' into $\mathbb{R}$, such that:

(a) *$\{Y_n\}$ and Y have distribution functions $\{F_n\}$ and F,*

(b) *$Y_n \xrightarrow{\text{a.s.}} Y$ as $n \to \infty$.*

Therefore, although X_n may fail to converge to X in any mode other than in distribution, there exists a sequence $\{Y_n\}$ such that Y_n is distributed identically to X_n for every n, which converges almost surely to a copy of X. The proof is elementary.

Proof. Let $\Omega' = (0, 1)$, let $\mathcal{F}'$ be the Borel σ-field generated by the intervals of Ω' (see the discussion at the end of Section 4.1), and let $\mathbb{P}'$ be the probability measure induced on $\mathcal{F}'$ by the requirement that, for any interval $I = (a, b) \subseteq \Omega'$, $\mathbb{P}'(I) = (b - a)$; $\mathbb{P}'$ is called *Lebesgue measure*. For $\omega \in \Omega'$, define

$$Y_n(\omega) = \inf\{x : \omega \leq F_n(x)\},$$
$$Y(\omega) = \inf\{x : \omega \leq F(x)\}.$$

Note that Y_n and Y are essentially the inverse functions of F_n and F since

$$\omega \leq F_n(x) \Leftrightarrow Y_n(\omega) \leq x, \tag{15}$$
$$\omega \leq F(x) \Leftrightarrow Y(\omega) \leq x.$$

It follows immediately that Y_n and Y satisfy (14a) since, for example, from (15)

$$\mathbb{P}'(Y \leq y) = \mathbb{P}'\big((0, F(y)]\big) = F(y).$$

To show (14b), proceed as follows. Given $\epsilon > 0$ and $\omega \in \Omega'$, pick a point x of continuity of F such that

$$Y(\omega) - \epsilon < x < Y(\omega).$$

We have by (15) that $F(x) < \omega$. However, $F_n(x) \to F(x)$ as $n \to \infty$ and so $F_n(x) < \omega$ for all large n, giving that

$$Y(\omega) - \epsilon < x < Y_n(\omega) \quad \text{for all large } n;$$

now let $n \to \infty$ and $\epsilon \downarrow 0$ to obtain

$$\liminf_{n\to\infty} Y_n(\omega) \geq Y(\omega) \quad \text{for all } \omega. \tag{16}$$

Finally, if $\omega < \omega' < 1$, pick a point x of continuity of F such that

$$Y(\omega') < x < Y(\omega') + \epsilon.$$

We have by (15) that $\omega < \omega' \leq F(x)$, and so $\omega < F_n(x)$ for all large n, giving that

$$Y_n(\omega) \leq x < Y(\omega') + \epsilon \quad \text{for all large } n;$$

now let $n \to \infty$ and $\epsilon \downarrow 0$ to obtain

$$\limsup_{n\to\infty} Y_n(\omega) \leq Y(\omega') \quad \text{whenever } \omega < \omega'. \tag{17}$$

Combine this with (16) to see that $Y_n(\omega) \to Y(\omega)$ for all points ω of continuity of Y. However, Y is monotone non-decreasing and so that set D of discontinuities of Y is countable; thus $\mathbb{P}'(D) = 0$ and the proof is complete. ■

We complete this section with two elementary applications of the representation theorem (14). The results in question are standard, but the usual classical proofs are tedious.

(18) Theorem. *If* $X_n \xrightarrow{\mathrm{D}} X$ *and* $g : \mathbb{R} \to \mathbb{R}$ *is continuous then* $g(X_n) \xrightarrow{\mathrm{D}} g(X)$.

Proof. Let $\{Y_n\}$ and Y be given as in (14). By the continuity of g,

$$\{\omega : g(Y_n(\omega)) \to g(Y(\omega))\} \supseteq \{\omega : Y_n(\omega) \to Y(\omega)\},$$

and so $g(Y_n) \xrightarrow{\text{a.s.}} g(Y)$ as $n \to \infty$. Therefore $g(Y_n) \xrightarrow{\mathrm{D}} g(Y)$; however, $g(Y_n)$ and $g(Y)$ have the same distributions as $g(X_n)$ and $g(X)$. ■

(19) Theorem. *The following three statements are equivalent.*

(a) $X_n \xrightarrow{\mathrm{D}} X$.

(b) $\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$ *for all bounded continuous functions* g.

(c) $\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$ *for all functions* g *of the form* $g(x) = f(x)I_{[a,b]}(x)$ *where* f *is continuous on* $[a, b]$ *and* a *and* b *are points of continuity of the distribution function of the random variable* X.

Condition (b) is usually taken as the definition of what is called *weak convergence*. It is not important in (c) that g be continuous on the *closed* interval $[a, b]$. The same proof is valid if g in part (c) is of the form $g(x) = f(x)I_{(a,b)}(x)$ where f is bounded and continuous on the open interval (a, b).

Proof. First we prove that (a) implies (b). Suppose that $X_n \xrightarrow{\mathrm{D}} X$ and g is bounded and continuous. By the Skorokhod representation theorem (14), there exist random variables $Y, Y_1, Y_2, \ldots$ having the same distributions as $X, X_1, X_2, \ldots$ and such that $Y_n \xrightarrow{\text{a.s.}} Y$. Therefore $g(Y_n) \xrightarrow{\text{a.s.}} g(Y)$ by the continuity of g, and furthermore the $g(Y_n)$ are uniformly bounded random variables. We apply the bounded convergence theorem (5.6.12) to deduce that $\mathbb{E}(g(Y_n)) \to \mathbb{E}(g(Y))$, and (b) follows since $\mathbb{E}(g(Y_n)) = \mathbb{E}(g(X_n))$ and $\mathbb{E}(g(Y)) = \mathbb{E}(g(X))$.

We write C for the set of points of continuity of F_X. Now F_X is monotone and has therefore at most countably many points of discontinuity; hence C^{c} is countable.

Suppose now that (b) holds. For (c), it suffices to prove that $\mathbb{E}(h(X_n)) \to \mathbb{E}(h(X))$ for all functions h of the form $h(x) = f(x)I_{(-\infty,b]}(x)$, where f is bounded and continuous, and $b \in C$; the general result follows by an exactly analogous argument. Suppose then that $h(x) = f(x)I_{(-\infty,b]}(x)$ as prescribed. The idea is to approximate to h by a continuous function. For $\delta > 0$, define the continuous functions h' and h'' by

$$h'(x) = \begin{cases} h(x) & \text{if } x \notin (b, b+\delta), \\ \left(1 + \dfrac{b-x}{\delta}\right) h(b) & \text{if } x \in (b, b+\delta), \end{cases}$$

$$h''(x) = \begin{cases} \left(1 + \dfrac{x-b}{\delta}\right) h(b) & \text{if } x \in (b-\delta, b), \\ \left(1 + \dfrac{b-x}{\delta}\right) h(b) & \text{if } x \in [b, b+\delta), \\ 0 & \text{otherwise.} \end{cases}$$

It may be helpful to draw a picture. Now

$$\big|\mathbb{E}\big(h(X_n) - h'(X_n)\big)\big| \le \big|\mathbb{E}(h''(X_n))\big|, \qquad \big|\mathbb{E}\big(h(X) - h'(X)\big)\big| \le \big|\mathbb{E}(h''(X))\big|$$

so that

$$\begin{aligned}\big|\mathbb{E}(h(X_n)) - \mathbb{E}(h(X))\big| &\le \big|\mathbb{E}(h''(X_n))\big| + \big|\mathbb{E}(h''(X))\big| + \big|\mathbb{E}(h'(X_n)) - \mathbb{E}(h'(X))\big| \\ &\to 2\big|\mathbb{E}(h''(X))\big| \quad \text{as} \quad n \to \infty\end{aligned}$$

by assumption (b). We now observe that

$$\big|\mathbb{E}(h''(X))\big| \le |h(b)|\mathbb{P}(b - \delta < X < b + \delta) \to 0 \quad \text{as} \quad \delta \downarrow 0,$$

by the assumption that $\mathbb{P}(X = b) = 0$. Hence (c) holds.

Suppose finally that (c) holds, and that b is such that $\mathbb{P}(X = b) = 0$. By considering the function $f(x) = 1$ for all x, we have that, if $a \in C$,

$$\begin{aligned}\textbf{(20)} \qquad \mathbb{P}(X_n \le b) \ge \mathbb{P}(a \le X_n \le b) &\to \mathbb{P}(a \le X \le b) \quad \text{as } n \to \infty \\ &\to \mathbb{P}(X \le b) \qquad \text{as } a \to -\infty \text{ through } C.\end{aligned}$$

A similar argument, but taking the limit in the other direction, yields for $b' \in C$

$$\begin{aligned}\textbf{(21)} \qquad \mathbb{P}(X_n \ge b') \ge \mathbb{P}(b' \le X_n \le c) &\qquad \text{if } c \ge b' \\ \to \mathbb{P}(b' \le X \le c) &\qquad \text{as } n \to \infty, \text{ if } c \in C \\ \to \mathbb{P}(X \ge b') &\qquad \text{as } c \to \infty \text{ through } C.\end{aligned}$$

It follows from (20) and (21) that, if $b, b' \in C$ and $b < b'$, then for any $\epsilon > 0$ there exists N such that

$$\mathbb{P}(X \le b) - \epsilon \le \mathbb{P}(X_n \le b) \le \mathbb{P}(X_n < b') \le \mathbb{P}(X < b') + \epsilon$$

for all $n \ge N$. Take the limits as $n \to \infty$ and $\epsilon \downarrow 0$, and $b' \downarrow b$ through C, in that order, to obtain that $\mathbb{P}(X_n \le b) \to \mathbb{P}(X \le b)$ as $n \to \infty$ if $b \in C$, the required result. ■

Exercises for Section 7.2

1. (a) Suppose $X_n \xrightarrow{r} X$ where $r \ge 1$. Show that $\mathbb{E}|X_n^r| \to \mathbb{E}|X^r|$.

(b) Suppose $X_n \xrightarrow{1} X$. Show that $\mathbb{E}(X_n) \to \mathbb{E}(X)$. Is the converse true?

(c) Suppose $X_n \xrightarrow{2} X$. Show that $\text{var}(X_n) \to \text{var}(X)$.

2. Dominated convergence. Suppose $|X_n| \le Z$ for all n, where $\mathbb{E}(Z) < \infty$. Prove that if $X_n \xrightarrow{\text{P}} X$ then $X_n \xrightarrow{1} X$.

3. Give a rigorous proof that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ for any pair X, Y of independent non-negative random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite means. [Hint: For $k \ge 0$, $n \ge 1$, define $X_n = k/n$ if $k/n \le X < (k+1)/n$, and similarly for Y_n. Show that X_n and Y_n are independent, and $X_n \le X$, and $Y_n \le Y$. Deduce that $\mathbb{E}X_n \to \mathbb{E}X$ and $\mathbb{E}Y_n \to \mathbb{E}Y$, and also $\mathbb{E}(X_nY_n) \to \mathbb{E}(XY)$.]

4. Show that convergence in distribution is equivalent to convergence with respect to the Lévy metric of Exercise (7.1.4).

5. (a) Suppose that $X_n \xrightarrow{\mathrm{D}} X$ and $Y_n \xrightarrow{\mathrm{P}} c$, where c is a constant. Show that $X_n Y_n \xrightarrow{\mathrm{D}} cX$, and that $X_n/Y_n \xrightarrow{\mathrm{D}} X/c$ if $c \neq 0$.

(b) Suppose that $X_n \xrightarrow{\mathrm{D}} 0$ and $Y_n \xrightarrow{\mathrm{P}} Y$, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be such that $g(x, y)$ is a continuous function of y for all x, and $g(x, y)$ is continuous at $x = 0$ for all y. Show that $g(X_n, Y_n) \xrightarrow{\mathrm{P}} g(0, Y)$.

[These results are sometimes referred to as 'Slutsky's theorem(s)'.]

6. Let $X_1, X_2, \ldots$ be random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Show that the set $A = \{\omega \in \Omega : \text{the sequence } X_n(\omega) \text{ converges}\}$ is an event (that is, lies in $\mathcal{F}$), and that there exists a random variable X (that is, an $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$) such that $X_n(\omega) \to X(\omega)$ for $\omega \in A$.

7. Let $\{X_n\}$ be a sequence of random variables, and let $\{c_n\}$ be a sequence of reals converging to the limit c. For convergence almost surely, in rth mean, in probability, and in distribution, show that the convergence of X_n to X entails the convergence of $c_n X_n$ to cX.

8. Let $\{X_n\}$ be a sequence of independent random variables which converges in probability to the limit X. Show that X is almost surely constant.

9. Convergence in total variation. The sequence of discrete random variables X_n, with mass functions f_n, is said to *converge in total variation* to X with mass function f if

$$\sum_x |f_n(x) - f(x)| \to 0 \quad \text{as} \quad n \to \infty.$$

Suppose $X_n \to X$ in total variation, and $u : \mathbb{R} \to \mathbb{R}$ is bounded. Show that $\mathbb{E}(u(X_n)) \to \mathbb{E}(u(X))$.

10. Let $\{X_r : r \geq 1\}$ be independent Poisson variables with respective parameters $\{\lambda_r : r \geq 1\}$. Show that $\sum_{r=1}^{\infty} X_r$ converges or diverges almost surely according as $\sum_{r=1}^{\infty} \lambda_r$ converges or diverges.

7.3 Some ancillary results

Next we shall develop some refinements of the methods of the last section; these will prove to be of great value later. There are two areas of interest. The first deals with inequalities and generalizes Markov's inequality, Lemma (7.2.7). The second deals with infinite families of events and the Borel–Cantelli lemmas; it is related to the result of Theorem (7.2.4c).

Markov's inequality is easily generalized.

(1) Theorem. *Let $h : \mathbb{R} \to [0, \infty)$ be a non-negative function. Then*

$$\mathbb{P}\big(h(X) \geq a\big) \leq \frac{\mathbb{E}(h(X))}{a} \quad \textit{for all } a > 0.$$

Proof. Denote by A the event $\{h(X) \geq a\}$, so that $h(X) \geq aI_A$. Take expectations to obtain the result. ■

We note some special cases of this.

(2) Example. Markov's inequality. Set $h(x) = |x|$. ●

(3) Example†. Chebyshov's inequality. Set $h(x) = x^2$ to obtain

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(X^2)}{a^2} \quad \text{if} \quad a > 0.$$

This inequality was also discovered by Bienaymé and others. ●

(4) Example. More generally, let $g : [0, \infty) \to [0, \infty)$ be a strictly increasing non-negative function, and set $h(x) = g(|x|)$ to obtain

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}\big(g(|X|)\big)}{g(a)} \quad \text{if} \quad a > 0. \qquad \bullet$$

Theorem (1) provides an upper bound for the probability $\mathbb{P}(h(X) \geq a)$. Lower bounds are harder to find in general, but pose no difficulty in the case when h is a uniformly bounded function.

(5) Theorem. *If $h : \mathbb{R} \to [0, M]$ is a non-negative function taking values bounded by some number M, then*

$$\mathbb{P}\big(h(X) \geq a\big) \geq \frac{\mathbb{E}(h(X)) - a}{M - a} \quad \textit{whenever} \quad 0 \leq a < M.$$

Proof. Let $A = \{h(X) \geq a\}$ as before and note that $h(X) \leq MI_A + aI_{A^c}$. ●

The reader is left to apply this result to the special cases (2), (3), and (4). This is an appropriate moment to note three other important inequalities. Let X and Y be random variables.

(6) Theorem. Hölder's inequality. *If $p, q > 1$ and $p^{-1} + q^{-1}$ then*

$$\mathbb{E}|XY| \leq (\mathbb{E}|X^p|)^{1/p}(\mathbb{E}|Y^q|)^{1/q}.$$

(7) Theorem. Minkowski's inequality. *If $p \geq 1$ then*

$$\big[\mathbb{E}(|X + Y|^p)\big]^{1/p} \leq (\mathbb{E}|X^p|)^{1/p} + (\mathbb{E}|Y^p|)^{1/p}.$$

Proof of (6) and (7). You did these for Problem (4.14.27). ■

(8) Theorem. $\mathbb{E}(|X + Y|^p) \leq C_p\big[\mathbb{E}|X^p| + \mathbb{E}|Y^p|\big]$ *where $p > 0$ and*

$$C_p = \begin{cases} 1 & \text{if } 0 < p \leq 1, \\ 2^{p-1} & \text{if } p > 1. \end{cases}$$

†Our transliteration of Чебышёв (Chebyshov) is at odds with common practice, but dispenses with the need for clairvoyance in pronunciation.

Proof. It is not difficult to show that $|x+y|^p \le C_p[|x|^p + |y|^p]$ for all $x, y \in \mathbb{R}$ and $p > 0$. Now complete the details. ■

Inequalities (6) and (7) assert that

$$\begin{aligned} \|XY\|_1 &\le \|X\|_p \|Y\|_q \qquad \text{if } p^{-1} + q^{-1} = 1, \\ \|X+Y\|_p &\le \|X\|_p + \|Y\|_p \qquad \text{if } p \ge 1, \end{aligned}$$

where $\|\cdot\|_p$ is the L_p norm $\|X\|_p = (\mathbb{E}|X^p|)^{1/p}$.

Here is an application of these inequalities. It is related to the fact that if $x_n \to x$ and $y_n \to y$ then $x_n + y_n \to x + y$.

(9) Theorem.

(a) *If* $X_n \xrightarrow{\text{a.s.}} X$ *and* $Y_n \xrightarrow{\text{a.s.}} Y$ *then* $X_n + Y_n \xrightarrow{\text{a.s.}} X + Y$.

(b) *If* $X_n \xrightarrow{r} X$ *and* $Y_n \xrightarrow{r} Y$ *then* $X_n + Y_n \xrightarrow{r} X + Y$.

(c) *If* $X_n \xrightarrow{\text{P}} X$ *and* $Y_n \xrightarrow{\text{P}} Y$ *then* $X_n + Y_n \xrightarrow{\text{P}} X + Y$.

(d) *It is not in general true that* $X_n + Y_n \xrightarrow{\text{D}} X + Y$ *whenever* $X_n \xrightarrow{\text{D}} X$ *and* $Y_n \xrightarrow{\text{D}} Y$.

Proof. *You* do it. You will need either (7) or (8) to prove part (b). ■

Theorem (7.2.4) contains a criterion for a sequence to converge almost surely. It is a special case of two very useful results called the 'Borel–Cantelli lemmas'. Let $A_1, A_2, \ldots$ be an infinite sequence of events from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We shall often be interested in finding out how many of the A_n occur. Recall (Problem (1.8.16)) that the event that infinitely many of the A_n occur, sometimes written $\{A_n \text{ infinitely often}\}$ or $\{A_n \text{ i.o.}\}$, satisfies

$$\{A_n \text{ i.o.}\} = \limsup_{n\to\infty} A_n = \bigcap_n \bigcup_{m=n}^{\infty} A_m.$$

(10) Theorem. Borel–Cantelli lemmas. *Let* $A = \bigcap_n \bigcup_{m=n}^{\infty} A_m$ *be the event that infinitely many of the* A_n *occur. Then*:

(a) $\mathbb{P}(A) = 0$ *if* $\sum_n \mathbb{P}(A_n) < \infty$,

(b) $\mathbb{P}(A) = 1$ *if* $\sum_n \mathbb{P}(A_n) = \infty$ *and* $A_1, A_2, \ldots$ *are independent events.*

It is easy to see that statement (b) is false if the assumption of independence is dropped. Just consider some event E with $0 < \mathbb{P}(E) < 1$ and define $A_n = E$ for all n. Then $A = E$ and $\mathbb{P}(A) = \mathbb{P}(E)$.

Proof.

(a) We have that $A \subseteq \bigcup_{m=n}^{\infty} A_m$ for all n, and so

$$\mathbb{P}(A) \le \sum_{m=n}^{\infty} \mathbb{P}(A_m) \to 0 \quad \text{as} \quad n \to \infty,$$

whenever $\sum_n \mathbb{P}(A_n) < \infty$.

(b) It is an easy *exercise* in set theory to check that

$$A^c = \bigcup_n \bigcap_{m=n}^{\infty} A_m^c.$$

However,

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^c\Big) &= \lim_{r\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{r} A_m^c\Big) && \text{by Lemma (1.3.5)} \\
&= \prod_{m=n}^{\infty} [1 - \mathbb{P}(A_m)] && \text{by independence} \\
&\le \prod_{m=n}^{\infty} \exp[-\mathbb{P}(A_m)] && \text{since } 1 - x \le e^{-x} \text{ if } x \ge 0 \\
&= \exp\Big(-\sum_{m=n}^{\infty} \mathbb{P}(A_m)\Big) = 0
\end{aligned}$$

whenever $\sum_n \mathbb{P}(A_n) = \infty$. Thus

$$\mathbb{P}(A^c) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^c\Big) = 0,$$

giving $\mathbb{P}(A) = 1$ as required. ■

(11) **Example. Markov chains.** Let $\{X_n\}$ be a Markov chain with $X_0 = i$ for some state i. Let $A_n = \{X_n = i\}$ be the event that the chain returns to i after n steps. State i is persistent if and only if $\mathbb{P}(A_n \text{ i.o.}) = 1$. By the first Borel–Cantelli lemma,

$$\mathbb{P}(A_n \text{ i.o.}) = 0 \quad \text{if} \quad \sum_n \mathbb{P}(A_n) < \infty$$

and it follows that i is transient if $\sum_n p_{ii}(n) < \infty$, which is part of an earlier result, Corollary (6.2.4). We cannot establish the converse by this method since the A_n are not independent. ●

If the events $A_1, A_2, \ldots$ of Theorem (10) are independent then $\mathbb{P}(A)$ equals either 0 or 1 depending on whether or not $\sum \mathbb{P}(A_n)$ converges. This is an example of a general theorem called a 'zero–one law'. There are many such results, of which the following is a simple example.

(12) **Theorem. Zero–one law.** *Let $A_1, A_2, \ldots$ be a collection of events, and let $\mathcal{A}$ be the smallest σ-field of subsets of Ω which contains all of them. If $A \in \mathcal{A}$ is an event which is independent of the finite collection $A_1, A_2, \ldots, A_n$ for each value of n, then*

$$\textit{either} \quad \mathbb{P}(A) = 0 \quad \textit{or} \quad \mathbb{P}(A) = 1.$$

Proof. Roughly speaking, the assertion that A belongs to $\mathcal{A}$ means that A is definable in terms of $A_1, A_2, \ldots$. Examples of such events include B_1, B_2, and B_3 defined by

$$B_1 = A_7 \setminus A_9, \quad B_2 = A_3 \cup A_6 \cup A_9 \cup \cdots, \quad B_3 = \bigcup_n \bigcap_{m=n}^{\infty} A_m.$$

A standard result of measure theory asserts that if $A \in \mathcal{A}$ then there exists a sequence of events $\{C_n\}$ such that

$$\textbf{(13)} \qquad C_n \in \mathcal{A}_n \quad \text{and} \quad \mathbb{P}(A \triangle C_n) \to 0 \quad \text{as } n \to \infty$$

where $\mathcal{A}_n$ is the smallest σ-field which contains the finite collection $A_1, A_2, \ldots, A_n$. But A is assumed independent of this collection, and so is independent of C_n for all n. From (13),

$$\textbf{(14)} \qquad \mathbb{P}(A \cap C_n) \to \mathbb{P}(A).$$

However, by independence,

$$\mathbb{P}(A \cap C_n) = \mathbb{P}(A)\mathbb{P}(C_n) \to \mathbb{P}(A)^2$$

which may be combined with (14) to give $\mathbb{P}(A) = \mathbb{P}(A)^2$, and so $\mathbb{P}(A)$ is 0 or 1. ■

Read on for another zero–one law. Let $X_1, X_2, \ldots$ be a collection of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any subcollection $\{X_i : i \in I\}$, write $\sigma(X_i : i \in I)$ for the smallest σ-field with respect to which each of the variables X_i $(i \in I)$ is measurable. This σ-field exists by the argument of Section 1.6. It contains events which are 'defined in terms of $\{X_i : i \in I\}$'. Let $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$. Then $\mathcal{H}_n \supseteq \mathcal{H}_{n+1} \supseteq \cdots$; write

$$\mathcal{H}_\infty = \bigcap_n \mathcal{H}_n.$$

$\mathcal{H}_\infty$ is called the *tail σ-field* of the X_n and contains events such as

$$\{X_n > 0 \text{ i.o.}\}, \quad \left\{\limsup_{n\to\infty} X_n = \infty\right\}, \quad \left\{\sum_n X_n \text{ converges}\right\},$$

the definitions of which need never refer to any finite subcollection $\{X_1, X_2, \ldots, X_n\}$. Events in $\mathcal{H}_\infty$ are called *tail events*.

(15) Theorem. Kolmogorov's zero–one law. *If $X_1, X_2, \ldots$ are independent variables then all events $H \in \mathcal{H}_\infty$ satisfy either $\mathbb{P}(H) = 0$ or $\mathbb{P}(H) = 1$.*

Such a σ-field $\mathcal{H}_\infty$ is called *trivial* since it contains only null events and their complements. You may try to prove this theorem using the techniques in the proof of (12); it is not difficult.

(16) Example. Let $X_1, X_2, \ldots$ be independent random variables and let

$$H_1 = \left\{\omega \in \Omega : \sum_n X_n(\omega) \text{ converges}\right\},$$

$$H_2 = \left\{\omega \in \Omega : \limsup_{n\to\infty} X_n(\omega) = \infty\right\}.$$

Each H_i has either probability 0 or probability 1. ●

We can associate many other random variables with the sequence $X_1, X_2, \ldots$; these include

$$Y_1 = \tfrac{1}{2}(X_3 + X_6), \quad Y_2 = \limsup_{n\to\infty} X_n, \quad Y_3 = Y_1 + Y_2.$$

We call such a variable Y a *tail function* if it is $\mathcal{H}_\infty$-measurable, where $\mathcal{H}_\infty$ is the tail σ-field of the X_n. Roughly speaking, Y is a tail function if its definition includes no essential reference to any finite subsequence $X_1, X_2, \ldots, X_n$. The random variables Y_1 and Y_3 are *not* tail functions; can you see why Y_2 *is* a tail function? More rigorously (see the discussion after Definition (2.1.3)) Y is a tail function if and only if

$$\{\omega \in \Omega : Y(\omega) \leq y\} \in \mathcal{H}_\infty \quad \text{for all } y \in \mathbb{R}.$$

Thus, if $\mathcal{H}_\infty$ is trivial then the distribution function $F_Y(y) = \mathbb{P}(Y \leq y)$ takes the values 0 and 1 only. Such a function is the distribution function of a random variable which is constant (see Example (2.1.7)), and we have shown the following useful result.

(17) Theorem. *Let Y be a tail function of the independent sequence $X_1, X_2, \ldots$. There exists k satisfying $-\infty \leq k \leq \infty$ such that $\mathbb{P}(Y = k) = 1$.*

Proof. Let $k = \inf\{y : \mathbb{P}(Y \leq y) = 1\}$, with the convention that the infimum of an empty set is $+\infty$. Then

$$\mathbb{P}(Y \leq y) = \begin{cases} 0 & \text{if } y < k, \\ 1 & \text{if } y \geq k. \end{cases}$$

■

(18) Example. Let $X_1, X_2, \ldots$ be independent variables with partial sums $S_n = \sum_{i=1}^n X_i$. Then

$$Z_1 = \liminf_{n\to\infty} \frac{1}{n} S_n, \quad Z_2 = \limsup_{n\to\infty} \frac{1}{n} S_n$$

are almost surely constant (but possibly infinite). To see this, note that if $m \leq n$ then

$$\frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{m} X_i + \frac{1}{n} \sum_{i=m+1}^{n} X_i = S_n(1) + S_n(2), \text{ say.}$$

However, $S_n(1) \to 0$ pointwise as $n \to \infty$, and so Z_1 and Z_2 depend in no way upon the values of $X_1, X_2, \ldots, X_m$. It follows that the event

$$\left\{ \frac{1}{n} S_n \text{ converges} \right\} = \{Z_1 = Z_2\}$$

has either probability 1 or probability 0. That is, $n^{-1} S_n$ converges either almost everywhere or almost nowhere; this was, of course, deducible from (15) since $\{Z_1 = Z_2\} \in \mathcal{H}_\infty$. ●

Exercises for Section 7.3

1. (a) Suppose that $X_n \xrightarrow{\mathrm{P}} X$. Show that $\{X_n\}$ is *Cauchy convergent in probability* in that, for all $\epsilon > 0$, $\mathbb{P}(|X_n - X_m| > \epsilon) \to 0$ as $n, m \to \infty$. In what sense is the converse true?

(b) Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables such that the pairs (X_i, X_j) and (Y_i, Y_j) have the same distributions for all i, j. If $X_n \xrightarrow{\mathrm{P}} X$, show that Y_n converges in probability to some limit Y having the same distribution as X.

2. Show that the probability that infinitely many of the events $\{A_n : n \geq 1\}$ occur satisfies $\mathbb{P}(A_n \text{ i.o.}) \geq \limsup_{n\to\infty} \mathbb{P}(A_n)$.

3. Let $\{S_n : n \geq 0\}$ be a simple random walk which moves to the right with probability p at each step, and suppose that $S_0 = 0$. Write $X_n = S_n - S_{n-1}$.

(a) Show that $\{S_n = 0 \text{ i.o.}\}$ is not a tail event of the sequence $\{X_n\}$.

(b) Show that $\mathbb{P}(S_n = 0 \text{ i.o.}) = 0$ if $p \neq \frac{1}{2}$.

(c) Let $T_n = S_n/\sqrt{n}$, and show that

$$\left\{\liminf_{n\to\infty} T_n \leq -x\right\} \cap \left\{\limsup_{n\to\infty} T_n \geq x\right\}$$

is a tail event of the sequence $\{X_n\}$, for all $x > 0$, and deduce directly that $\mathbb{P}(S_n = 0 \text{ i.o.}) = 1$ if $p = \frac{1}{2}$.

4. Hewitt–Savage zero–one law. Let $X_1, X_2, \ldots$ be independent identically distributed random variables. The event A, defined in terms of the X_n, is called *exchangeable* if A is invariant under finite permutations of the coordinates, which is to say that its indicator function I_A satisfies $I_A(X_1, X_2, \ldots, X_n, \ldots) = I_A(X_{i_1}, X_{i_2}, \ldots, X_{i_n}, X_{n+1}, \ldots)$ for all $n \geq 1$ and all permutations $(i_1, i_2, \ldots, i_n)$ of $(1, 2, \ldots, n)$. Show that all exchangeable events A are such that either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

5. Returning to the simple random walk S of Exercise (3), show that $\{S_n = 0 \text{ i.o.}\}$ is an exchangeable event with respect to the steps of the walk, and deduce from the Hewitt–Savage zero–one law that it has probability either 0 or 1.

6. Weierstrass's approximation theorem. Let $f : [0, 1] \to \mathbb{R}$ be a continuous function, and let S_n be a random variable having the binomial distribution with parameters n and x. Using the formula $\mathbb{E}(Z) = \mathbb{E}(ZI_A) + \mathbb{E}(ZI_{A^c})$ with $Z = f(x) - f(n^{-1}S_n)$ and $A = \{|n^{-1}S_n - x| > \delta\}$, show that

$$\lim_{n\to\infty} \sup_{0\leq x\leq 1} \left| f(x) - \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k} \right| = 0.$$

You have proved Weierstrass's approximation theorem, which states that every continuous function on $[0, 1]$ may be approximated by a polynomial uniformly over the interval.

7. Complete convergence. A sequence $X_1, X_2, \ldots$ of random variables is said to be *completely convergent* to X if

$$\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty \qquad \text{for all } \epsilon > 0.$$

Show that, for sequences of independent variables, complete convergence is equivalent to a.s. convergence. Find a sequence of (dependent) random variables which converges a.s. but not completely.

8. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with common mean μ and finite variance. Show that

$$\binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} X_i X_j \xrightarrow{\mathrm{P}} \mu^2 \qquad \text{as } n \to \infty.$$

9. Let $\{X_n : n \geq 1\}$ be independent and exponentially distributed with parameter 1. Show that

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} = 1\right) = 1.$$

10. Let $\{X_n : n \geq 1\}$ be independent $N(0, 1)$ random variables. Show that:

(a) $\mathbb{P}\left(\limsup_{n\to\infty} \dfrac{|X_n|}{\sqrt{\log n}} = \sqrt{2}\right) = 1,$

(b) $\mathbb{P}(X_n > a_n \text{ i.o.}) = \begin{cases} 0 & \text{if } \sum_n \mathbb{P}(X_1 > a_n) < \infty, \\ 1 & \text{if } \sum_n \mathbb{P}(X_1 > a_n) = \infty. \end{cases}$

11. Construct an example to show that the convergence in distribution of X_n to X does not imply the convergence of the unique medians of the sequence X_n.

12. (i) Let $\{X_r : r \geq 1\}$ be independent, non-negative and identically distributed with infinite mean. Show that $\limsup_{r\to\infty} X_r/r = \infty$ almost surely.
(ii) Let $\{X_r\}$ be a stationary Markov chain on the positive integers with transition probabilities

$$p_{jk} = \begin{cases} \dfrac{j}{j+2} & \text{if } k = j+1, \\ \dfrac{2}{j+2} & \text{if } k = 1. \end{cases}$$

(a) Find the stationary distribution of the chain, and show that it has infinite mean.
(b) Show that $\limsup_{r\to\infty} X_r/r \leq 1$ almost surely.

13. Let $\{X_r : 1 \leq r \leq n\}$ be independent and identically distributed with mean μ and finite variance σ^2. Let $\overline{X} = n^{-1}\sum_{r=1}^n X_r$. Show that

$$\sum_{r=1}^n (X_r - \mu) \Big/ \sqrt{\sum_{r=1}^n (X_r - \overline{X})^2}$$

converges in distribution to the $N(0, 1)$ distribution as $n \to \infty$.

7.4 Laws of large numbers

Let $\{X_n\}$ be a sequence of random variables with partial sums $S_n = \sum_{i=1}^n X_i$. We are interested in the asymptotic behaviour of S_n as $n \to \infty$; this long-term behaviour depends crucially upon the sequence $\{X_i\}$. The general problem may be described as follows. Under what conditions does the following convergence occur?

$$\frac{S_n}{b_n} - a_n \to S \quad \text{as} \quad n \to \infty \tag{1}$$

where $a = \{a_n\}$ and $b = \{b_n\}$ are sequences of real numbers, S is a random variable, and the convergence takes place in some mode to be specified.

(2) Example. Let $X_1, X_2, \ldots$ be independent identically distributed variables with mean μ and variance σ^2. By Theorems (5.10.2) and (5.10.4), we have that

$$\frac{S_n}{n} \xrightarrow{\text{D}} \mu \quad \text{and} \quad \frac{S_n}{\sigma\sqrt{n}} - \frac{\mu\sqrt{n}}{\sigma} \xrightarrow{\text{D}} N(0, 1).$$

There may not be a *unique* collection a, b, S such that (1) occurs. ●

The convergence problem (1) can often be simplified by setting $a_n = 0$ for all n, whenever the X_i have finite means. Just rewrite the problem in terms of $X_i' = X_i - \mathbb{E}X_i$ and $S_n' = S_n - \mathbb{E}S_n$.

The general theory of relations such as (1) is well established and extensive. We shall restrict our attention here to a small but significant part of the theory when the X_i are independent and identically distributed random variables. Suppose for the moment that this is true. We saw in Example (2) that (at least) two types of convergence may be established for such sequences, so long as they have finite second moments. The law of large numbers admits stronger forms than that given in (2). For example, notice that $n^{-1}S_n$ converges in distribution to a constant limit, and use Theorem (7.2.4) to see that $n^{-1}S_n$ converges in probability also. Perhaps we can strengthen this further to include convergence in rth mean, for some r, or almost sure convergence. Indeed, this turns out to be possible when suitable conditions are imposed on the common distribution of the X_i. We shall not use the method of characteristic functions of Chapter 5, preferring to approach the problem more directly in the spirit of Section 7.2.

We shall say that the sequence $\{X_n\}$ obeys the 'weak law of large numbers' if there exists a constant μ such that $n^{-1}S_n \xrightarrow{\text{P}} \mu$. If the stronger result $n^{-1}S_n \xrightarrow{\text{a.s.}} \mu$ holds, then we call it the 'strong law of large numbers'. We seek sufficient, and if possible necessary, conditions on the common distribution of the X_i for the weak and strong laws to hold. As the title suggests, the weak law is implied by the strong law, since convergence in probability is implied by almost sure convergence. A sufficient condition for the strong law is given by the following theorem.

(3) Theorem. *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbb{E}(X_1^2) < \infty$ and $\mathbb{E}(X_1) = \mu$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \textit{almost surely and in mean square.}$$

This strong law holds whenever the X_i have finite second moment. The proof of mean square convergence is very easy; almost sure convergence is harder to demonstrate (but see Problem (7.11.6) for an easy proof of almost sure convergence subject to the stronger condition that $\mathbb{E}(X_1^4) < \infty$).

Proof. To show mean square convergence, calculate

$$\begin{aligned}
\mathbb{E}\left(\left(\frac{1}{n}S_n - \mu\right)^2\right) &= \mathbb{E}\left(\frac{1}{n^2}(S_n - \mathbb{E}S_n)^2\right) = \frac{1}{n^2}\operatorname{var}\left(\sum_1^n X_i\right) \\
&= \frac{1}{n^2}\sum_1^n \operatorname{var}(X_i) && \text{by independence and Theorem (3.3.11)} \\
&= \frac{1}{n}\operatorname{var}(X_1) \to 0 && \text{as } n \to \infty,
\end{aligned}$$

since $\operatorname{var}(X_1) < \infty$ by virtue of the assumption that $\mathbb{E}(X_1^2) < \infty$.

Next we show almost sure convergence. We saw in Theorem (7.2.13) that there necessarily exists a subsequence $n_1, n_2, \ldots$ along which $n^{-1}S_n$ converges to μ almost surely; we can find such a subsequence explicitly. Write $n_i = i^2$ and use Chebyshov's inequality, Example (7.3.3), to find that

$$\mathbb{P}\left(\frac{1}{i^2}|S_{i^2} - i^2\mu| > \epsilon\right) \leq \frac{\operatorname{var}(S_{i^2})}{i^4\epsilon^2} = \frac{\operatorname{var}(X_1)}{i^2\epsilon^2}.$$

Sum over i and use Theorem (7.2.4c) to find that

$$\frac{1}{i^2}S_{i^2} \xrightarrow{\text{a.s.}} \mu \quad \text{as} \quad i \to \infty. \tag{4}$$

We need to fill in the gaps in this limit process. Suppose for the moment that the X_i are *non-negative*. Then $\{S_n\}$ is monotonic non-decreasing, and so

$$S_{i^2} \le S_n \le S_{(i+1)^2} \quad \text{if} \quad i^2 \le n \le (i+1)^2.$$

Divide by n to find that

$$\frac{1}{(i+1)^2}S_{i^2} \le \frac{1}{n}S_n \le \frac{1}{i^2}S_{(i+1)^2} \quad \text{if} \quad i^2 \le n \le (i+1)^2;$$

now let $n \to \infty$ and use (4), remembering that $i^2/(i+1)^2 \to 1$ as $i \to \infty$, to deduce that

$$\frac{1}{n}S_n \xrightarrow{\text{a.s.}} \mu \quad \text{as} \quad n \to \infty \tag{5}$$

as required, whenever the X_i are non-negative. Finally we lift the non-negativity condition. For general X_i, define random variables X_n^+, X_n^- by

$$X_n^+(\omega) = \max\{X_n(\omega), 0\}, \quad X_n^-(\omega) = -\min\{X_n(\omega), 0\};$$

then X_n^+ and X_n^- are non-negative and

$$X_n = X_n^+ - X_n^-, \quad \mathbb{E}(X_n) = \mathbb{E}(X_n^+) - \mathbb{E}(X_n^-).$$

Furthermore, $X_n^+ \le |X_n|$ and $X_n^- \le |X_n|$, so that $\mathbb{E}((X_1^+)^2) < \infty$ and $\mathbb{E}((X_1^-)^2) < \infty$. Now apply (5) to the sequences $\{X_n^+\}$ and $\{X_n^-\}$ to find, by Theorem (7.3.9a), that

$$\begin{aligned}\frac{1}{n}S_n &= \frac{1}{n}\left(\sum_1^n X_i^+ - \sum_1^n X_i^-\right) \\ &\xrightarrow{\text{a.s.}} \mathbb{E}(X_1^+) - \mathbb{E}(X_1^-) = \mathbb{E}(X_1) \quad \text{as} \quad n \to \infty.\end{aligned}$$

■

Is the result of Theorem (3) as sharp as possible? It is not difficult to see that the condition $\mathbb{E}(X_1^2) < \infty$ is both necessary and sufficient for mean square convergence to hold. For almost sure convergence the weaker condition that

$$\mathbb{E}|X_1| < \infty \tag{6}$$

will turn out to be necessary and sufficient, but the proof of this is slightly more difficult and is deferred until the next section. There exist sequences which satisfy the weak law but not the strong law. Indeed, the characteristic function technique (see Section 5.10) can be used to prove the following necessary and sufficient condition for the weak law. We offer no proof, but see Laha and Rohatgi (1979, p. 320), Feller (1971, p. 565), and Problem (7.11.15).

(7) Theorem. *The independent identically distributed sequence $\{X_n\}$, with common distribution function F, satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mu$$

for some constant μ, if and only if one of the following conditions (8) *or* (9) *holds*:

(8) $$n\mathbb{P}(|X_1| > n) \to 0 \quad \text{and} \quad \int_{[-n,n]} x\,dF \to \mu \quad \text{as } n \to \infty,$$

(9) *the characteristic function $\phi(t)$ of the X_j is differentiable at $t = 0$ and $\phi'(0) = i\mu$.*

Of course, the integral in (8) can be rewritten as

$$\int_{[-n,n]} x\,dF = \mathbb{E}\big(X_1 \,\big|\, |X_1| \le n\big)\mathbb{P}(|X_1| \le n) = \mathbb{E}(X_1 I_{\{|X_1| \le n\}}).$$

Thus, a sequence satisfies the weak law but not the strong law whenever (8) holds without (6); as an example of this, suppose the X_j are symmetric (in that X_1 and $-X_1$ have the same distribution) but their common distribution function F satisfies

$$1 - F(x) \sim \frac{1}{x \log x} \quad \text{as} \quad x \to \infty.$$

Some distributions fail even to satisfy (8).

(10) Example. Let the X_j have the Cauchy distribution with density function

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

Then the first part of (8) is violated. Indeed, the characteristic function of $U_n = n^{-1}S_n$ is

$$\phi_{U_n}(t) = \phi_{X_1}\left(\frac{t}{n}\right) \cdots \phi_{X_n}\left(\frac{t}{n}\right) = \left[\exp\left(-\frac{|t|}{n}\right)\right]^n = e^{-|t|}$$

and so U_n itself has the Cauchy distribution for all values of n. In particular, (1) holds with $b_n = n$, $a_n = 0$, where S is Cauchy, and the convergence is in distribution. ●

Exercises for Section 7.4

1. Let $X_2, X_3, \ldots$ be independent random variables such that

$$\mathbb{P}(X_n = n) = \mathbb{P}(X_n = -n) = \frac{1}{2n \log n}, \quad \mathbb{P}(X_n = 0) = 1 - \frac{1}{n \log n}.$$

Show that this sequence obeys the weak law but not the strong law, in the sense that $n^{-1}\sum_1^n X_i$ converges to 0 in probability but not almost surely.

2. Construct a sequence $\{X_r : r \ge 1\}$ of independent random variables with zero mean such that $n^{-1}\sum_{r=1}^n X_r \to -\infty$ almost surely, as $n \to \infty$.

3. Let N be a spatial Poisson process with constant intensity λ in $\mathbb{R}^d$, where $d \ge 2$. Let S be the ball of radius r centred at zero. Show that $N(S)/|S| \to \lambda$ almost surely as $r \to \infty$, where $|S|$ is the volume of the ball.

7.5 The strong law

This section is devoted to the proof of the strong law of large numbers.

(1) Theorem. Strong law of large numbers. *Let $X_1, X_2, \dots$ be independent identically distributed random variables. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \text{ almost surely, as } n \to \infty,$$

for some constant μ, if and only if $\mathbb{E}|X_1| < \infty$. In this case $\mu = \mathbb{E}X_1$.

The traditional proof of this theorem is long and difficult, and proceeds by a generalization of Chebyshov's inequality. We avoid that here, and give a relatively elementary proof which is an adaptation of the method used to prove Theorem (7.4.3). We make use of the technique of *truncation*, used earlier in the proof of the large deviation theorem (5.11.4).

Proof. Suppose first that the X_i are *non-negative* random variables with $\mathbb{E}|X_1| = \mathbb{E}(X_1) < \infty$, and write $\mu = \mathbb{E}(X_1)$. We 'truncate' the X_n to obtain a new sequence $\{Y_n\}$ given by

$$Y_n = X_n I_{\{X_n < n\}} = \begin{cases} X_n & \text{if } X_n < n, \\ 0 & \text{if } X_n \geq n. \end{cases} \tag{2}$$

Note that

$$\sum_n \mathbb{P}(X_n \neq Y_n) = \sum_n \mathbb{P}(X_n \geq n) \leq \mathbb{E}(X_1) < \infty$$

by the result of Problem (4.14.3). Of course, $\mathbb{P}(X_n \geq n) = \mathbb{P}(X_1 \geq n)$ since the X_i are identically distributed. By the first Borel–Cantelli lemma (7.3.10a),

$$\mathbb{P}\big(X_n \neq Y_n \text{ for infinitely many values of } n\big) = 0,$$

and so

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad n \to \infty; \tag{3}$$

thus it will suffice to show that

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{\text{a.s.}} \mu \quad \text{as} \quad n \to \infty. \tag{4}$$

We shall need the following elementary observation. If $\alpha > 1$ and $\beta_k = \lfloor \alpha^k \rfloor$, the integer part of α^k, then there exists $A > 0$ such that

$$\sum_{k=m}^{\infty} \frac{1}{\beta_k^2} \leq \frac{A}{\beta_m^2} \quad \text{for} \quad m \geq 1. \tag{5}$$

This holds because, for large m, the convergent series on the left side is 'nearly' geometric with first term β_m^{-2}. Note also that

$$\frac{\beta_{k+1}}{\beta_k} \to \alpha \quad \text{as} \quad k \to \infty. \tag{6}$$

Write $S_n' = \sum_{i=1}^n Y_i$. For $\alpha > 1, \epsilon > 0$, use Chebyshov's inequality to find that

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{1}{\beta_n}|S'_{\beta_n} - \mathbb{E}(S'_{\beta_n})| > \epsilon\right) &\le \frac{1}{\epsilon^2}\sum_{n=1}^{\infty}\frac{1}{\beta_n^2}\operatorname{var}(S'_{\beta_n}) \\ &= \frac{1}{\epsilon^2}\sum_{n=1}^{\infty}\frac{1}{\beta_n^2}\sum_{i=1}^{\beta_n}\operatorname{var}(Y_i) \quad \text{by independence} \\ &\le \frac{A}{\epsilon^2}\sum_{i=1}^{\infty}\frac{1}{i^2}\mathbb{E}(Y_i^2) \end{aligned} \tag{7}$$

by changing the order of summation and using (5).

Let $B_{ij} = \{j - 1 \le X_i < j\}$, and note that $\mathbb{P}(B_{ij}) = \mathbb{P}(B_{1j})$. Now

$$\begin{aligned} \sum_{i=1}^{\infty}\frac{1}{i^2}\mathbb{E}(Y_i^2) &= \sum_{i=1}^{\infty}\frac{1}{i^2}\sum_{j=1}^{i}\mathbb{E}(Y_i^2 I_{B_{ij}}) \quad \text{by (2)} \\ &\le \sum_{i=1}^{\infty}\frac{1}{i^2}\sum_{j=1}^{i} j^2\mathbb{P}(B_{ij}) \\ &\le \sum_{j=1}^{\infty} j^2\mathbb{P}(B_{1j})\frac{2}{j} \le 2\big[\mathbb{E}(X_1) + 1\big] < \infty. \end{aligned} \tag{8}$$

Combine (7) and (8) and use Theorem (7.2.4c) to deduce that

$$\frac{1}{\beta_n}\big[S'_{\beta_n} - \mathbb{E}(S'_{\beta_n})\big] \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad n \to \infty. \tag{9}$$

Also,

$$\mathbb{E}(Y_n) = \mathbb{E}(X_n I_{\{X_n < n\}}) = \mathbb{E}(X_1 I_{\{X_1 < n\}}) \to \mathbb{E}(X_1) = \mu$$

as $n \to \infty$, by monotone convergence (5.6.12). Thus

$$\frac{1}{\beta_n}\mathbb{E}(S'_{\beta_n}) = \frac{1}{\beta_n}\sum_{i=1}^{\beta_n}\mathbb{E}(Y_i) \to \mu \quad \text{as} \quad n \to \infty$$

(remember the hint in the proof of Corollary (6.4.22)), yielding from (9) that

$$\frac{1}{\beta_n}S'_{\beta_n} \xrightarrow{\text{a.s.}} \mu \quad \text{as} \quad n \to \infty; \tag{10}$$

this is a partial demonstration of (4). In order to fill in the gaps, use the fact that the Y_i are non-negative, implying that the sequence $\{S'_n\}$ is monotonic non-decreasing, to deduce that

$$\textbf{(11)} \qquad \frac{1}{\beta_{n+1}} S'_{\beta_n} \le \frac{1}{m} S'_m \le \frac{1}{\beta_n} S'_{\beta_{n+1}} \quad \text{if} \quad \beta_n \le m \le \beta_{n+1}.$$

Let $m \to \infty$ in (11) and remember (6) to find that

$$\textbf{(12)} \qquad \alpha^{-1}\mu \le \liminf_{m\to\infty} \frac{1}{m} S'_m \le \limsup_{m\to\infty} \frac{1}{m} S'_m \le \alpha\mu \quad \text{almost surely.}$$

This holds for all $\alpha > 1$; let $\alpha \downarrow 1$ to obtain (4), and deduce by (3) that

$$\textbf{(13)} \qquad \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu \quad \text{as} \quad n \to \infty$$

whenever the X_i are non-negative. Now proceed exactly as in the proof of Theorem (7.4.3) in order to lift the non-negativity condition. Note that we have proved the main part of the theorem without using the full strength of the independence assumption; we have used only the fact that the X_i are *pairwise* independent.

In order to prove the converse, suppose that $n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu$. Then $n^{-1}X_n \xrightarrow{\text{a.s.}} 0$ by the theory of convergent real series, and the second Borel–Cantelli lemma (7.3.10b) gives

$$\sum_n \mathbb{P}(|X_n| \ge n) < \infty,$$

since the divergence of this sum would imply that $\mathbb{P}(n^{-1}|X_n| \ge 1 \text{ i.o.}) = 1$ (only here do we use the full assumption of independence). By Problem (4.14.3),

$$\mathbb{E}|X_1| \le 1 + \sum_{n=1}^{\infty} \mathbb{P}(|X_1| \ge n) = 1 + \sum_{n=1}^{\infty} \mathbb{P}(|X_n| \ge n),$$

and hence $\mathbb{E}|X_1| < \infty$, which completes the proof of the theorem. ■

Exercises for Section 7.5

1. Entropy. The interval $[0, 1]$ is partitioned into n disjoint sub-intervals with lengths $p_1, p_2, \ldots, p_n$, and the *entropy* of this partition is defined to be

$$h = -\sum_{i=1}^{n} p_i \log p_i.$$

Let $X_1, X_2, \ldots$ be independent random variables having the uniform distribution on $[0, 1]$, and let $Z_m(i)$ be the number of the $X_1, X_2, \ldots, X_m$ which lie in the ith interval of the partition above. Show that

$$R_m = \prod_{i=1}^{n} p_i^{Z_m(i)}$$

satisfies $m^{-1} \log R_m \to -h$ almost surely as $m \to \infty$.

2. Recurrent events. Catastrophes occur at the times $T_1, T_2, \ldots$ where $T_i = X_1 + X_2 + \cdots + X_i$ and the X_i are independent identically distributed positive random variables. Let $N(t) = \max\{n : T_n \le t\}$ be the number of catastrophes which have occurred by time t. Prove that if $\mathbb{E}(X_1) < \infty$ then $N(t) \to \infty$ and $N(t)/t \to 1/\mathbb{E}(X_1)$ as $t \to \infty$, almost surely.

3. Random walk. Let $X_1, X_2, \ldots$ be independent identically distributed random variables taking values in the integers $\mathbb{Z}$ and having a finite mean. Show that the Markov chain $S = \{S_n\}$ given by $S_n = \sum_1^n X_i$ is transient if $\mathbb{E}(X_1) \ne 0$.

7.6 The law of the iterated logarithm

Let $S_n = X_1 + X_2 + \cdots + X_n$ be the partial sum of independent identically distributed variables, as usual, and suppose further that $\mathbb{E}(X_i) = 0$ and $\text{var}(X_i) = 1$ for all i. To date, we have two results about the growth rate of $\{S_n\}$.

Law of large numbers: $\frac{1}{n}S_n \to 0$ a.s. and in mean square.

Central limit theorem: $\frac{1}{\sqrt{n}}S_n \xrightarrow{\text{D}} N(0, 1)$.

Thus the sequence $U_n = S_n/\sqrt{n}$ enjoys a random fluctuation which is asymptotically regularly distributed. Apart from this long-term trend towards the normal distribution, the sequence $\{U_n\}$ may suffer some large but rare fluctuations. The law of the iterated logarithm is an extraordinary result which tells us exactly how large these fluctuations are. First note that, in the language of Section 7.3,

$$U = \limsup_{n\to\infty} \frac{U_n}{\sqrt{2\log\log n}}$$

is a tail function of the sequence of the X_i. The zero–one law, Theorem (7.3.17), tells us that there exists a number k, possibly infinite, such that $\mathbb{P}(U = k) = 1$. The next theorem asserts that $k = 1$.

(1) Theorem. Law of the iterated logarithm. *If $X_1, X_2, \ldots$ are independent identically distributed random variables with mean 0 and variance 1 then*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1\right) = 1.$$

The proof is long and difficult and is omitted (but see the discussion in Billingsley (1995) or Laha and Rohatgi (1979)). The theorem amounts to the assertion that

$$A_n = \left\{S_n \geq c\sqrt{2n\log\log n}\right\}$$

occurs for infinitely many values of n if $c < 1$ and for only finitely many values of n if $c > 1$, with probability 1. It is an immediate corollary of (1) that

$$\mathbb{P}\left(\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1\right) = 1;$$

just apply (1) to the sequence $-X_1, -X_2, \ldots$.

Exercise for Section 7.6

1. A function $\phi(x)$ is said to belong to the 'upper class' if, in the notation of this section, $\mathbb{P}(S_n > \phi(n)\sqrt{n}\ \text{i.o.}) = 0$. A consequence of the law of the iterated logarithm is that $\sqrt{\alpha\log\log x}$ is in the upper class for all $\alpha > 2$. Use the first Borel–Cantelli lemma to prove the much weaker fact that $\phi(x) = \sqrt{\alpha\log x}$ is in the upper class for all $\alpha > 2$, in the special case when the X_i are independent $N(0, 1)$ variables.

7.7 Martingales

Many probabilists specialize in limit theorems, and much of applied probability is devoted to finding such results. The accumulated literature is vast and the techniques multifarious. One of the most useful skills for establishing such results is that of martingale divination, because the convergence of martingales is guaranteed.

(1) Example. It is appropriate to discuss an example of the use of the word 'martingale' which pertains to gambling, a favourite source of probabilistic illustrations. We are all familiar with the following gambling strategy. A gambler has a large fortune. He wagers £1 on an evens bet. If he loses then he wagers £2 on the next play. If he loses on the nth play then he wagers $£2^n$ on the next. Each sum is calculated so that his inevitable ultimate win will cover his lost stakes and profit him by £1. This strategy is called a 'martingale'. Nowadays casinos do not allow its use, and croupiers have instructions to refuse the bets of those who are seen to practise it. Thackeray's advice was to avoid its use at all costs, and his reasoning may have had something to do with the following calculation. Suppose the gambler wins for the first time at the Nth play. N is a random variable with mass function

$$\mathbb{P}(N = n) = (\tfrac{1}{2})^n$$

and so $\mathbb{P}(N < \infty) = 1$; the gambler is almost surely guaranteed a win in the long run. However, by this time he will have lost an amount $£L$ with mean value

$$\mathbb{E}(L) = \sum_{n=1}^{\infty} (\tfrac{1}{2})^n (1 + 2 + \cdots + 2^{n-2}) = \infty.$$

He must be prepared to lose a lot of money! And so, of course, must the proprietor of the casino.

The perils of playing the martingale are illustrated by the following two excerpts from the memoirs of G. Casanova recalling his stay in Venice in 1754 (Casanova 1922, Chapter 7).

> Playing the martingale, continually doubling my stake, I won every day during the rest of the carnival. I was fortunate enough never to lose the sixth card, and if I had lost it, I should have been without money to play, for I had 2000 sequins on that card. I congratulated myself on having increased the fortune of my dear mistress.

However, some days later:

> I still played the martingale, but with such bad luck that I was soon left without a sequin. As I shared my property with my mistress, I was obliged to tell her of my losses, and at her request sold all her diamonds, losing what I got for them; she had now only 500 sequins. There was no more talk of her escaping from the convent, for we had nothing to live on.

Shortly after these events, Casanova was imprisoned by the authorities, until he escaped to organize a lottery for the benefit of both himself and the French treasury in Paris. Before it became merely a spangle, the sequin was an Italian gold coin. ●

In the spirit of this diversion, suppose a gambler wagers repeatedly with an initial capital S_0, and let S_n be his capital after n plays. We shall think of $S_0, S_1, \ldots$ as a sequence of dependent random variables. Before his $(n + 1)$th wager the gambler knows the numerical values of $S_0, S_1, \ldots, S_n$, but can only guess at the future $S_{n+1}, \ldots$. If the game is fair then, conditional

upon the past information, he will expect no change in his present capital on average. That is to say†,

$$\mathbb{E}(S_{n+1} \mid S_0, S_1, \ldots, S_n) = S_n. \tag{2}$$

Most casinos need to pay at least their overheads, and will find a way of changing this equation to

$$\mathbb{E}(S_{n+1} \mid S_0, S_1, \ldots, S_n) \leq S_n.$$

The gambler is fortunate indeed if this inequality is reversed. Sequences satisfying (2) are called 'martingales', and they have very special and well-studied properties of convergence. They may be discovered within many probabilistic models, and their general theory may be used to establish limit theorems. We shall now abandon the gambling example, and refer disappointed readers to *How to gamble if you must* by L. Dubins and L. Savage, where they may find an account of the gamblers' ruin theorem.

(3) Definition. A sequence $\{S_n : n \geq 1\}$ is a **martingale** with respect to the sequence $\{X_n : n \geq 1\}$ if, for all $n \geq 1$:

(a) $\mathbb{E}|S_n| < \infty$,

(b) $\mathbb{E}(S_{n+1} \mid X_1, X_2, \ldots, X_n) = S_n$.

Equation (2) shows that the sequence of gambler's fortunes is a martingale with respect to itself. The extra generality, introduced by the sequence $\{X_n\}$ in (3), is useful for martingales which arise in the following way. A specified sequence $\{X_n\}$ of random variables, such as a Markov chain, may itself *not* be a martingale. However, it is often possible to find some function ϕ such that $\{S_n = \phi(X_n) : n \geq 1\}$ *is* a martingale. In this case, the martingale property (2) becomes the assertion that, given the values of $X_1, X_2, \ldots, X_n$, the mean value of $S_{n+1} = \phi(X_{n+1})$ is just $S_n = \phi(X_n)$; that is,

$$\mathbb{E}(S_{n+1} \mid X_1, \ldots, X_n) = S_n. \tag{4}$$

Of course, this condition is without meaning unless S_n is some function, say ϕ_n, of $X_1, \ldots, X_n$ (that is, $S_n = \phi_n(X_1, \ldots, X_n)$) since the conditional expectation in (4) is itself a function of $X_1, \ldots, X_n$. We shall often omit reference to the underlying sequence $\{X_n\}$, asserting merely that $\{S_n\}$ is a martingale.

(5) Example. Branching processes, two martingales. Let Z_n be the size of the nth generation of a branching process, with $Z_0 = 1$. Recall that the probability η that the process ultimately becomes extinct is the smallest non-negative root of the equation $s = G(s)$, where G is the probability generating function of Z_1. There are (at least) two martingales associated with the process. First, conditional on $Z_n = z_n$, Z_{n+1} is the sum of z_n independent family sizes, and so

$$\mathbb{E}(Z_{n+1} \mid Z_n = z_n) = z_n \mu$$

†Such conditional expectations appear often in this section. Make sure you understand their meanings. This one is the mean value of S_{n+1}, calculated as though $S_0, \ldots, S_n$ were already known. Clearly this mean value depends on $S_0, \ldots, S_n$; so it is a *function* of $S_0, \ldots, S_n$. Assertion (2) is that it has the value S_n. Any detailed account of conditional expectations would probe into the guts of measure theory. We shall avoid that here, but describe some important properties at the end of this section and in Section 7.9.

where $\mu = G'(1)$ is the mean family size. Thus, by the Markov property,

$$\mathbb{E}(Z_{n+1} \mid Z_1, Z_2, \ldots, Z_n) = Z_n\mu.$$

Now define $W_n = Z_n/\mathbb{E}(Z_n)$ and remember that $\mathbb{E}(Z_n) = \mu^n$ to obtain

$$\mathbb{E}(W_{n+1} \mid Z_1, \ldots, Z_n) = W_n,$$

and so $\{W_n\}$ is a martingale (with respect to $\{Z_n\}$). It is not the only martingale which arises from the branching process. Let $V_n = \eta^{Z_n}$ where η is the probability of ultimate extinction. Surprisingly perhaps, $\{V_n\}$ is a martingale also, as the following indicates. Write $Z_{n+1} = X_1 + X_2 + \cdots + X_{Z_n}$ in terms of the family sizes of the members of the nth generation to obtain

$$\begin{aligned}\mathbb{E}(V_{n+1} \mid Z_1, \ldots, Z_n) &= \mathbb{E}\big(\eta^{(X_1+\cdots+X_{Z_n})} \,\big|\, Z_1, \ldots, Z_n\big) \\ &= \prod_{i=1}^{Z_n} \mathbb{E}(\eta^{X_i} \mid Z_1, \ldots, Z_n) \quad \text{by independence} \\ &= \prod_{i=1}^{Z_n} \mathbb{E}(\eta^{X_i}) = \prod_{i=1}^{Z_n} G(\eta) = \eta^{Z_n} = V_n,\end{aligned}$$

since $\eta = G(\eta)$. These facts are very significant in the study of the long-term behaviour of the branching process. ●

(6) Example. Let $X_1, X_2, \ldots$ be independent variables with zero means. We claim that the sequence of partial sums $S_n = X_1 + X_2 + \cdots + X_n$ is a martingale (with respect to $\{X_n\}$). For,

$$\begin{aligned}\mathbb{E}(S_{n+1} \mid X_1, \ldots, X_n) &= \mathbb{E}(S_n + X_{n+1} \mid X_1, \ldots, X_n) \\ &= \mathbb{E}(S_n \mid X_1, \ldots, X_n) + \mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) \\ &= S_n + 0, \quad \text{by independence.}\end{aligned}$$ ●

(7) Example. Markov chains. Let $X_0, X_1, \ldots$ be a discrete-time Markov chain taking values in some countable state space S with transition matrix $\mathbf{P}$. Suppose that $\psi : S \to \mathbb{R}$ is a bounded function which satisfies

$$\sum_{j \in S} p_{ij}\psi(j) = \psi(i) \quad \text{for all } i \in S. \tag{8}$$

We claim that $S_n = \psi(X_n)$ constitutes a martingale (with respect to $\{X_n\}$). For,

$$\begin{aligned}\mathbb{E}(S_{n+1} \mid X_1, \ldots, X_n) &= \mathbb{E}\big(\psi(X_{n+1}) \,\big|\, X_1, \ldots, X_n\big) \\ &= \mathbb{E}\big(\psi(X_{n+1}) \,\big|\, X_n\big) \quad \text{by the Markov property} \\ &= \sum_{j \in S} p_{X_n, j}\psi(j) \\ &= \psi(X_n) = S_n \quad \text{by (8).}\end{aligned}$$ ●

(9) Example. Let $X_1, X_2, \ldots$ be independent variables with zero means, finite variances, and partial sums $S_n = \sum_{i=1}^n X_i$. Define

$$T_n = S_n^2 = \left(\sum_{i=1}^n X_i\right)^2.$$

Then

$$\begin{aligned}\mathbb{E}(T_{n+1} \mid X_1, \ldots, X_n) &= \mathbb{E}\big(S_n^2 + 2S_nX_{n+1} + X_{n+1}^2 \,\big|\, X_1, \ldots, X_n\big) \\ &= T_n + 2\mathbb{E}(X_{n+1})\mathbb{E}(S_n \mid X_1, \ldots, X_n) + \mathbb{E}(X_{n+1}^2) \\ &\qquad\qquad \text{by independence} \\ &= T_n + \mathbb{E}(X_{n+1}^2) \geq T_n.\end{aligned}$$

Thus $\{T_n\}$ is not a martingale, since it only satisfies (4) with $\geq$ in place of $=$; it is called a 'submartingale', and has properties similar to those of a martingale. ●

These examples show that martingales are all around us. They are extremely useful because, subject to a condition on their moments, they always converge; this is 'Doob's convergence theorem' and is the main result of the next section. Martingales are explored in considerable detail in Chapter 12.

Finally, here are some properties of conditional expectation. You need not read them now, but may refer back to them when necessary. Recall that the conditional expectation of X given Y is defined by

$$\mathbb{E}(X \mid Y) = \psi(Y) \quad \text{where} \quad \psi(y) = \mathbb{E}(X \mid Y = y)$$

is the mean of the conditional distribution of X given that $Y = y$. Most of the conditional expectations in this chapter take the form $\mathbb{E}(X \mid \mathbf{Y})$, the mean value of X conditional on the values of the variables in the random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. We stress that $\mathbb{E}(X \mid \mathbf{Y})$ is a function of $\mathbf{Y}$ alone. Expressions such as '$\mathbb{E}(X \mid \mathbf{Y}) = Z$' should sometimes be qualified by 'almost surely'; we generally omit this qualification.

(10) Lemma.

(a) $\mathbb{E}(X_1 + X_2 \mid \mathbf{Y}) = \mathbb{E}(X_1 \mid \mathbf{Y}) + \mathbb{E}(X_2 \mid \mathbf{Y})$.

(b) $\mathbb{E}(Xg(\mathbf{Y}) \mid \mathbf{Y}) = g(\mathbf{Y})\mathbb{E}(X \mid \mathbf{Y})$ *for (measurable) functions* $g : \mathbb{R}^n \to \mathbb{R}$.

(c) $\mathbb{E}(X \mid h(\mathbf{Y})) = \mathbb{E}(X \mid \mathbf{Y})$ *if* $h : \mathbb{R}^n \to \mathbb{R}^n$ *is one–one.*

Sketch proof.

(a) This depends on the linearity of expectation only.

(b) $\mathbb{E}(Xg(\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}) = g(\mathbf{y})\mathbb{E}(X \mid \mathbf{Y} = \mathbf{y})$.

(c) Roughly speaking, knowledge of $\mathbf{Y}$ is interchangeable with knowledge of $h(\mathbf{Y})$, in that

$$\mathbf{Y}(\omega) = \mathbf{y} \text{ if and only if } h(\mathbf{Y}(\omega)) = h(\mathbf{y}), \quad \text{for any } \omega \in \Omega. \qquad \blacksquare$$

(11) Lemma. Tower property. $\mathbb{E}\big[\mathbb{E}(X \mid \mathbf{Y}_1, \mathbf{Y}_2) \,\big|\, \mathbf{Y}_1\big] = \mathbb{E}(X \mid \mathbf{Y}_1)$.

Proof. Just write down these expectations as integrals involving conditional distributions to see that the result holds. It is a more general version of Problem (4.14.29). ■

Sometimes we consider the mean value $\mathbb{E}(X \mid A)$ of a random variable X conditional upon the occurrence of some event A having strictly positive probability. This is just the mean of the corresponding distribution function $F_{X|A}(x) = \mathbb{P}(X \leq x \mid A)$. We can think of $\mathbb{E}(X \mid A)$ as a constant random variable with domain $A \subseteq \Omega$; it is undefined at points $\omega \in A^c$. The following result is an application of Lemma (1.4.4).

(12) Lemma. *If* $\{B_i : 1 \leq i \leq n\}$ *is a partition of* A *then*

$$\mathbb{E}(X \mid A)\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{E}(X \mid B_i)\mathbb{P}(B_i).$$

You may like the following proof:

$$\mathbb{E}(XI_A) = \mathbb{E}\Big(X \sum_i I_{B_i}\Big) = \sum_i \mathbb{E}(XI_{B_i}).$$

Sometimes we consider mixtures of these two types of conditional expectation. These are of the form $\mathbb{E}(X \mid \mathbf{Y}, A)$ where $X, Y_1, \ldots, Y_n$ are random variables and A is an event. Such quantities are defined in the obvious way and have the usual properties. For example, (11) becomes

$$\mathbb{E}(X \mid A) = \mathbb{E}\big[\mathbb{E}(X \mid \mathbf{Y}, A) \,\big|\, A\big]. \tag{13}$$

We shall make some use of the following fact soon. If, in (13), A is an event which is defined in terms of the Y_i (such as $A = \{Y_1 \leq 1\}$ or $A = \{|Y_2Y_3 - Y_4| > 2\}$) then it is not difficult to see that

$$\mathbb{E}\big[\mathbb{E}(X \mid \mathbf{Y}) \,\big|\, A\big] = \mathbb{E}\big[\mathbb{E}(X \mid \mathbf{Y}, A) \,\big|\, A\big]; \tag{14}$$

just note that evaluating the random variable $\mathbb{E}(X \mid \mathbf{Y}, A)$ at a point $\omega \in \Omega$ yields

$$\mathbb{E}(X \mid \mathbf{Y}, A)(\omega) \begin{cases} = \mathbb{E}(X \mid \mathbf{Y})(\omega) & \text{if } \omega \in A \\ \text{is undefined} & \text{if } \omega \notin A. \end{cases}$$

The sequences $\{S_n\}$ of this section satisfy

$$\mathbb{E}|S_n| < \infty, \quad \mathbb{E}(S_{n+1} \mid X_1, \ldots, X_n) = S_n. \tag{15}$$

(16) Lemma. *If* $\{S_n\}$ *satisfies* (15) *then*:

(a) $\mathbb{E}(S_{m+n} \mid X_1, \ldots, X_m) = S_m$ *for all* $m, n \geq 1$,

(b) $\mathbb{E}(S_n) = \mathbb{E}(S_1)$ *for all* n.

Proof.

(a) Use (11) with $X = S_{m+n}$, $\mathbf{Y}_1 = (X_1, \ldots, X_m)$, and $\mathbf{Y}_2 = (X_{m+1}, \ldots, X_{m+n-1})$ to obtain

$$\begin{aligned} \mathbb{E}(S_{m+n} \mid X_1, \ldots, X_m) &= \mathbb{E}\big[\mathbb{E}(S_{m+n} \mid X_1, \ldots, X_{m+n-1}) \,\big|\, X_1, \ldots, X_m\big] \\ &= \mathbb{E}(S_{m+n-1} \mid X_1, \ldots, X_m) \end{aligned}$$

and iterate to obtain the result.

(b) $\mathbb{E}(S_n) = \mathbb{E}\big(\mathbb{E}(S_n \mid X_1)\big) = \mathbb{E}(S_1)$ by (a). ■

For a more satisfactory account of conditional expectation, see Section 7.9.

Exercises for Section 7.7

1. Let $X_1, X_2, \ldots$ be random variables such that the partial sums $S_n = X_1 + X_2 + \cdots + X_n$ determine a martingale. Show that $\mathbb{E}(X_i X_j) = 0$ if $i \neq j$.

2. Let Z_n be the size of the nth generation of a branching process with immigration, in which the family sizes have mean μ ($\neq 1$) and the mean number of immigrants in each generation is m. Suppose that $\mathbb{E}(Z_0) < \infty$, and show that

$$S_n = \mu^{-n}\left\{Z_n - m\left(\frac{1-\mu^n}{1-\mu}\right)\right\}$$

is a martingale with respect to a suitable sequence of random variables.

3. Let $X_0, X_1, X_2, \ldots$ be a sequence of random variables with finite means and satisfying $\mathbb{E}(X_{n+1} \mid X_0, X_1, \ldots, X_n) = aX_n + bX_{n-1}$ for $n \geq 1$, where $0 < a, b < 1$ and $a + b = 1$. Find a value of α for which $S_n = \alpha X_n + X_{n-1}$, $n \geq 1$, defines a martingale with respect to the sequence X.

4. Let X_n be the net profit to the gambler of betting a unit stake on the nth play in a casino; the X_n may be dependent, but the game is fair in the sense that $\mathbb{E}(X_{n+1} \mid X_1, X_2, \ldots, X_n) = 0$ for all n. The gambler stakes Y on the first play, and thereafter stakes $f_n(X_1, X_2, \ldots, X_n)$ on the $(n+1)$th play, where $f_1, f_2, \ldots$ are given functions. Show that her profit after n plays is

$$S_n = \sum_{i=1}^{n} X_i f_{i-1}(X_1, X_2, \ldots, X_{i-1}),$$

where $f_0 = Y$. Show further that the sequence $S = \{S_n\}$ satisfies the martingale condition $\mathbb{E}(S_{n+1} \mid X_1, X_2, \ldots, X_n) = S_n$, $n \geq 1$, if Y is assumed to be known throughout.

7.8 Martingale convergence theorem

This section is devoted to the proof and subsequent applications of the following theorem. It receives a section to itself by virtue of its wealth of applications.

(1) Theorem. *If $\{S_n\}$ is a martingale with $\mathbb{E}(S_n^2) < M < \infty$ for some M and all n, then there exists a random variable S such that S_n converges to S almost surely and in mean square.*

This result has a more general version which, amongst other things,

(i) deals with submartingales,

(ii) imposes weaker moment conditions,

(iii) explores convergence in mean also,

but the proof of this is more difficult. On the other hand, the proof of (1) is within our grasp, and is only slightly more difficult than the proof of the strong law for independent sequences, Theorem (7.5.1); it mimics the traditional proof of the strong law and begins with a generalization of Chebyshov's inequality. We return to the theory of martingales in much greater generality in Chapter 12.

(2) Theorem. Doob–Kolmogorov inequality. *If $\{S_n\}$ is a martingale with respect to $\{X_n\}$ then*

$$\mathbb{P}\left(\max_{1\leq i\leq n} |S_i| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}(S_n^2) \quad \textit{whenever} \quad \epsilon > 0.$$

Proof of (2). Let $A_0 = \Omega$, $A_k = \{|S_i| < \epsilon \text{ for all } i \le k\}$, and let $B_k = A_{k-1} \cap \{|S_k| \ge \epsilon\}$ be the event that $|S_i| \ge \epsilon$ for the first time when $i = k$. Then

$$A_k \cup \Big(\bigcup_{i=1}^{k} B_i\Big) = \Omega.$$

Therefore

$$\mathbb{E}(S_n^2) = \sum_{i=1}^{n} \mathbb{E}(S_n^2 I_{B_i}) + \mathbb{E}(S_n^2 I_{A_n}) \ge \sum_{i=1}^{n} \mathbb{E}(S_n^2 I_{B_i}). \tag{3}$$

However,

$$\begin{aligned}\mathbb{E}(S_n^2 I_{B_i}) &= \mathbb{E}\big((S_n - S_i + S_i)^2 I_{B_i}\big) \\ &= \mathbb{E}\big((S_n - S_i)^2 I_{B_i}\big) + 2\mathbb{E}\big((S_n - S_i) S_i I_{B_i}\big) + \mathbb{E}(S_i^2 I_{B_i}) \\ &= \alpha + \beta + \gamma, \text{ say.}\end{aligned}$$

Note that $\alpha \ge 0$ and $\gamma \ge \epsilon^2 \mathbb{P}(B_i)$, because $|S_i| \ge \epsilon$ if B_i occurs. To deal with β, note that

$$\begin{aligned}\mathbb{E}\big((S_n - S_i) S_i I_{B_i}\big) &= \mathbb{E}\big[S_i I_{B_i} \mathbb{E}(S_n - S_i \mid X_1, \ldots, X_i)\big] && \text{by Lemma (7.7.10b)} \\ &= 0 && \text{by Lemma (7.7.16a),}\end{aligned}$$

since B_i concerns $X_1, \ldots, X_i$ only, by the discussion after (7.7.4). Thus (3) becomes

$$\mathbb{E}(S_n^2) \ge \sum_{i=1}^{n} \epsilon^2 \mathbb{P}(B_i) = \epsilon^2 \mathbb{P}\Big(\max_{1 \le i \le n} |S_i| \ge \epsilon\Big)$$

and the result is shown. ■

Proof of (1). First note that S_m and $(S_{m+n} - S_m)$ are uncorrelated whenever $m, n \ge 1$, since

$$\mathbb{E}\big(S_m(S_{m+n} - S_m)\big) = \mathbb{E}\big[S_m \mathbb{E}(S_{m+n} - S_m \mid X_1, \ldots, X_m)\big] = 0$$

by Lemma (7.7.16). Thus

$$\mathbb{E}(S_{m+n}^2) = \mathbb{E}(S_m^2) + \mathbb{E}\big((S_{m+n} - S_m)^2\big). \tag{4}$$

It follows that $\{\mathbb{E}(S_n^2)\}$ is a non-decreasing sequence, which is bounded above, by the assumption in (1); hence we may suppose that the constant M is chosen such that

$$\mathbb{E}(S_n^2) \uparrow M \quad \text{as} \quad n \to \infty.$$

We shall show that the sequence $\{S_n(\omega) : n \ge 1\}$ is almost-surely Cauchy convergent (see the notes on Cauchy convergence after Example (7.2.2)). Let $C = \{\omega \in \Omega : \{S_n(\omega)\}$

is Cauchy convergent}. For $\omega \in C$, the sequence $S_n(\omega)$ converges as $n \to \infty$ to some limit $S(\omega)$, and we shall show that $\mathbb{P}(C) = 1$. Note that

$$C = \Big\{\forall \epsilon > 0,\ \exists m \text{ such that } |S_{m+i} - S_{m+j}| < \epsilon \text{ for all } i, j \geq 1\Big\}.$$

By the triangle inequality

$$|S_{m+i} - S_{m+j}| \leq |S_{m+i} - S_m| + |S_{m+j} - S_m|,$$

so that

$$\begin{aligned} C &= \Big\{\forall \epsilon > 0,\ \exists m \text{ such that } |S_{m+i} - S_m| < \epsilon \text{ for all } i \geq 1\Big\} \\ &= \bigcap_{\epsilon > 0} \bigcup_m \big\{|S_{m+i} - S_m| < \epsilon \text{ for all } i \geq 1\big\}. \end{aligned}$$

The complement of C may be expressed as

$$C^{\mathrm{c}} = \bigcup_{\epsilon > 0} \bigcap_m \big\{|S_{m+i} - S_m| \geq \epsilon \text{ for some } i \geq 1\big\} = \bigcup_{\epsilon > 0} \bigcap_m A_m(\epsilon)$$

where $A_m(\epsilon) = \{|S_{m+i} - S_m| \geq \epsilon \text{ for some } i \geq 1\}$. Now $A_m(\epsilon) \subseteq A_m(\epsilon')$ if $\epsilon \geq \epsilon'$, so that

$$\mathbb{P}(C^{\mathrm{c}}) = \lim_{\epsilon \downarrow 0} \mathbb{P}\Big(\bigcap_m A_m(\epsilon)\Big) \leq \lim_{\epsilon \downarrow 0} \lim_{m \to \infty} \mathbb{P}(A_m(\epsilon)).$$

In order to prove that $\mathbb{P}(C^{\mathrm{c}}) = 0$ as required, it suffices to show that $\mathbb{P}(A_m(\epsilon)) \to 0$ as $m \to \infty$ for all $\epsilon > 0$. To this end we shall use the Doob–Kolmogorov inequality.

For a given choice of m, define the sequence $Y = \{Y_n : n \geq 1\}$ by $Y_n = S_{m+n} - S_m$. It may be checked that Y is a martingale with respect to itself:

$$\begin{aligned} \mathbb{E}(Y_{n+1} \mid Y_1, \ldots, Y_n) &= \mathbb{E}\big[\mathbb{E}(Y_{n+1} \mid X_1, \ldots, X_{m+n}) \,\big|\, Y_1, \ldots, Y_n\big] \\ &= \mathbb{E}(Y_n \mid Y_1, \ldots, Y_n) = Y_n \end{aligned}$$

by Lemma (7.7.11) and the martingale property. We apply the Doob–Kolmogorov inequality (2) to this martingale to find that

$$\mathbb{P}\big(|S_{m+i} - S_m| \geq \epsilon \text{ for some } 1 \leq i \leq n\big) \leq \frac{1}{\epsilon^2}\mathbb{E}\big((S_{m+n} - S_m)^2\big).$$

Letting $n \to \infty$ and using (4) we obtain

$$\mathbb{P}(A_m(\epsilon)) \leq \frac{1}{\epsilon^2}\big(M - \mathbb{E}(S_m^2)\big),$$

and hence $\mathbb{P}(A_m(\epsilon)) \to 0$ as $m \to \infty$ as required for almost-sure convergence. We have proved that there exists a random variable S such that $S_n \xrightarrow{\text{a.s.}} S$.

It remains only to prove convergence of S_n to S in mean square. For this we need Fatou's lemma (5.6.13). It is the case that

$$\begin{aligned}\mathbb{E}\big((S_n - S)^2\big) &= \mathbb{E}\Big(\liminf_{m\to\infty}(S_n - S_m)^2\Big) \le \liminf_{m\to\infty}\mathbb{E}\big((S_n - S_m)^2\big) \\ &= M - \mathbb{E}(S_n^2) \to 0 \quad \text{as} \quad n \to \infty,\end{aligned}$$

and the proof is finished. ■

Here are some applications of the martingale convergence theorem.

(5) Example. Branching processes. Recall Example (7.7.5). By Lemma (5.4.2), $W_n = Z_n/\mathbb{E}(Z_n)$ has second moment

$$\mathbb{E}(W_n^2) = 1 + \frac{\sigma^2(1 - \mu^{-n})}{\mu(\mu - 1)} \quad \text{if} \quad \mu \neq 1$$

where $\sigma^2 = \operatorname{var}(Z_1)$. Thus, if $\mu \neq 1$, there exists a random variable W such that $W_n \xrightarrow{\text{a.s.}} W$, and so $W_n \xrightarrow{\text{D}} W$ also; their characteristic functions satisfy $\phi_{W_n}(t) \to \phi_W(t)$ by Theorem (5.9.5). This makes the discussion at the end of Section 5.4 fully rigorous, and we can rewrite equation (5.4.6) as

$$\phi_W(\mu t) = G(\phi_W(t)).$$ ●

(6) Example. Markov chains. Suppose that the chain $X_0, X_1, \ldots$ of Example (7.7.7) is irreducible and persistent, and let ψ be a bounded function mapping S into $\mathbb{R}$ which satisfies equation (7.7.8). Then the sequence $\{S_n\}$, given by $S_n = \psi(X_n)$, is a martingale and satisfies the condition $\mathbb{E}(S_n^2) \le M$ for some M, by the boundedness of ψ. For any state i, the event $\{X_n = i\}$ occurs for infinitely many values of n with probability 1. However, $\{S_n = \psi(i)\} \supseteq \{X_n = i\}$ and so

$$S_n \xrightarrow{\text{a.s.}} \psi(i) \quad \text{for all } i,$$

which is clearly impossible unless $\psi(i)$ is the same for all i. We have shown that any bounded solution of equation (7.7.8) is constant. ●

(7) Example. Genetic model. Recall Example (6.1.11), which dealt with gene frequencies in the evolution of a population. We encountered there a Markov chain $X_0, X_1, \ldots$ taking values in $\{0, 1, \ldots, N\}$ with transition probabilities given by

$$\textbf{(8)} \qquad p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i) = \binom{N}{j}\left(\frac{i}{N}\right)^j\left(1 - \frac{i}{N}\right)^{N-j}.$$

Then

$$\begin{aligned}\mathbb{E}(X_{n+1} \mid X_0, \ldots, X_n) &= \mathbb{E}(X_{n+1} \mid X_n) \qquad \text{by the Markov property} \\ &= \sum_j j p_{X_n, j} = X_n\end{aligned}$$

by (8). Thus $X_0, X_1, \ldots$ is a martingale. Also, let Y_n be defined by $Y_n = X_n(N - X_n)$, and suppose that $N > 1$. Then

$$\mathbb{E}(Y_{n+1} \mid X_0, \ldots, X_n) = \mathbb{E}(Y_{n+1} \mid X_n) \quad \text{by the Markov property}$$

and

$$\mathbb{E}(Y_{n+1} \mid X_n = i) = \sum_j j(N-j)p_{ij} = i(N-i)(1 - N^{-1})$$

by (8). Thus

$$\mathbb{E}(Y_{n+1} \mid X_0, \ldots, X_n) = Y_n(1 - N^{-1}), \tag{9}$$

and we see that $\{Y_n\}$ is not itself a martingale. However, set $S_n = Y_n/(1 - N^{-1})^n$ to obtain from (9) that

$$\mathbb{E}(S_{n+1} \mid X_0, \ldots, X_n) = S_n;$$

we deduce that $\{S_n\}$ is a martingale.

The martingale $\{X_n\}$ has uniformly bounded second moments, and so there exists an X such that $X_n \xrightarrow{\text{a.s.}} X$. Unlike the previous example, this chain is not irreducible. In fact, 0 and N are absorbing states, and X takes these values only. Can you find the probability $\mathbb{P}(X = 0)$ that the chain is ultimately absorbed at 0? The results of the next section will help you with this.

Finally, what happens when we apply the convergence theorem to $\{S_n\}$? ●

Exercises for Section 7.8

1. Kolmogorov's inequality. Let $X_1, X_2, \ldots$ be independent random variables with zero means and finite variances, and let $S_n = X_1 + X_2 + \cdots + X_n$. Use the Doob–Kolmogorov inequality to show that

$$\mathbb{P}\left(\max_{1 \le j \le n} |S_j| > \epsilon\right) \le \frac{1}{\epsilon^2} \sum_{j=1}^{n} \operatorname{var}(X_j) \qquad \text{for } \epsilon > 0.$$

2. Let $X_1, X_2, \ldots$ be independent random variables such that $\sum_n n^{-2} \operatorname{var}(X_n) < \infty$. Use Kolmogorov's inequality to prove that

$$\sum_{i=1}^{n} \frac{X_i - \mathbb{E}(X_i)}{i} \xrightarrow{\text{a.s.}} Y \qquad \text{as } n \to \infty,$$

for some finite random variable Y, and deduce that

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \xrightarrow{\text{a.s.}} 0 \qquad \text{as } n \to \infty.$$

(You may find Kronecker's lemma to be useful: if (a_n) and (b_n) are real sequences with $b_n \uparrow \infty$ and $\sum_i a_i/b_i < \infty$, then $b_n^{-1} \sum_{i=1}^{n} a_i \to 0$ as $n \to \infty$.)

3. Let S be a martingale with respect to X, such that $\mathbb{E}(S_n^2) < K < \infty$ for some $K \in \mathbb{R}$. Suppose that $\operatorname{var}(S_n) \to 0$ as $n \to \infty$, and prove that $S = \lim_{n \to \infty} S_n$ exists and is constant almost surely.

7.9 Prediction and conditional expectation

Probability theory is not merely an intellectual pursuit, but provides also a framework for estimation and prediction. Practical men and women often need to make decisions based on quantities which are not easily measurable, either because they lie in the future or because of some intrinsic inaccessibility; in doing so they usually make use of some current or feasible observation. Economic examples are commonplace (business trends, inflation rates, and so on); other examples include weather prediction, the climate in prehistoric times, the state of the core of a nuclear reactor, the cause of a disease in an individual or a population, or the paths of celestial bodies. This last problem has the distinction of being amongst the first to be tackled by mathematicians using a modern approach to probability.

At its least complicated, a question of prediction or estimation involves an unknown or unobserved random variable Y, about which we are provided with the value of some (observable) random variable X. The problem is to deduce information about the value of Y from a knowledge of the value of X. Thus we seek a function $h(X)$ which is (in some sense) close to Y; we write $\widehat{Y} = h(X)$ and call $\widehat{Y}$ an 'estimator' of Y. As we saw in Section 7.1, there are many different ways in which two random variables may be said to be close to one another—pointwise, in rth mean, in probability, and so on. A particular way of especial convenience is to work with the norm given by

$$\|U\|_2 = \sqrt{\mathbb{E}(U^2)}, \tag{1}$$

so that the distance between two random variables U and V is

$$\|U - V\|_2 = \sqrt{\mathbb{E}\{(U - V)^2\}}; \tag{2}$$

The norm $\|\cdot\|_2$ is often called the L_2 norm, and the corresponding notion of convergence is of course convergence in mean square:

$$\|U_n - U\|_2 \to 0 \quad \text{if and only if} \quad U_n \xrightarrow{\text{m.s.}} U. \tag{3}$$

This norm is a special case of the 'L_p norm' given by $\|X\|_p = \{\mathbb{E}|X^p|\}^{1/p}$ where $p \geq 1$.

We recall that $\|\cdot\|_2$ satisfies the *triangle inequality*:

$$\|U + V\|_2 \leq \|U\|_2 + \|V\|_2. \tag{4}$$

With this notation, we make the following definition.

(5) **Definition.** Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(Y^2) < \infty$. The **minimum mean-squared-error predictor** (or **best predictor**) of Y given X is the function $\widehat{Y} = h(X)$ of X for which $\|Y - \widehat{Y}\|_2$ is a minimum.

We shall commonly use the term 'best predictor' in this context; the word 'best' is only shorthand, and should not be interpreted literally.

Let H be the set of all functions of X having finite second moment:

$$H = \{h(X) : h \text{ maps } \mathbb{R} \text{ to } \mathbb{R},\ \mathbb{E}(h(X)^2) < \infty\}. \tag{6}$$

The best (or minimum mean-squared-error) predictor of Y is (if it exists) a random variable $\widehat{Y}$ belonging to H such that $\mathbb{E}((Y - \widehat{Y})^2) \leq \mathbb{E}((Y - Z)^2)$ for all $Z \in H$. Does there exist

such a $\widehat{Y}$? The answer is yes, and furthermore there is (essentially) a unique such $\widehat{Y}$ in H. In proving this we shall make use of two properties of H, that it is a linear space, and that it is closed (with respect to the norm $\|\cdot\|_2$); that is to say, for $Z_1, Z_2, \ldots \in H$ and $a_1, a_2, \ldots \in \mathbb{R}$,

$$(7) \qquad a_1 Z_1 + a_2 Z_2 + \cdots + a_n Z_n \in H, \quad \text{and}$$

$$(8) \qquad \text{if } \|Z_m - Z_n\|_2 \to 0 \text{ as } m, n \to \infty, \text{ there exists } Z \in H \text{ such that } Z_n \xrightarrow{\text{m.s.}} Z.$$

(See Exercise (7.9.6a).) More generally, we call a set H of random variables a *closed linear space* (with respect to $\|\cdot\|_2$) if $\|X\|_2 < \infty$ for all $X \in H$, and H satisfies (7) and (8).

(9) Theorem. *Let H be a closed linear space (with respect to $\|\cdot\|_2$) of random variables. Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite variance. There exists a random variable $\widehat{Y}$ in H such that*

$$(10) \qquad \|Y - \widehat{Y}\|_2 \leq \|Y - Z\|_2 \quad \textit{for all } Z \in H,$$

and which is unique in the sense that $\mathbb{P}(\widehat{Y} = \overline{Y}) = 1$ for any $\overline{Y} \in H$ with $\|Y - \overline{Y}\|_2 = \|Y - \widehat{Y}\|_2$.

Proof. Let $d = \inf\{\|Y - Z\|_2 : Z \in H\}$, and find a sequence $Z_1, Z_2, \ldots$ in H such that $\lim_{n\to\infty} \|Y - Z_n\|_2 = d$. Now, for any $A, B \in H$, the 'parallelogram rule' holds†:

$$(11) \qquad \|A - B\|_2^2 = 2\Big[\|Y - A\|_2^2 - 2\big\|Y - \tfrac{1}{2}(A + B)\big\|_2^2 + \|Y - B\|_2^2\Big];$$

to show this, just expand the right side. Note that $\frac{1}{2}(A + B) \in H$ since H is a linear space. Setting $A = Z_n$, $B = Z_m$, we obtain using the definition of d that

$$\|Z_m - Z_n\|_2^2 \leq 2\Big(\|Y - Z_m\|_2^2 - 2d + \|Y - Z_n\|_2^2\Big) \to 0 \quad \text{as} \quad m, n \to \infty.$$

Therefore $\|Z_m - Z_n\|_2 \to 0$, so that there exists $\widehat{Y} \in H$ such that $Z_n \xrightarrow{\text{m.s.}} \widehat{Y}$; it is here that we use the fact (8) that H is closed. It follows by the triangle inequality (4) that

$$\|Y - \widehat{Y}\|_2 \leq \|Y - Z_n\|_2 + \|Z_n - \widehat{Y}\|_2 \to d \quad \text{as} \quad n \to \infty,$$

so that $\widehat{Y}$ satisfies (10).

Finally, suppose that $\overline{Y} \in H$ satisfies $\|Y - \overline{Y}\|_2 = d$. Apply (11) with $A = \overline{Y}$, $B = \widehat{Y}$, to obtain

$$\|\overline{Y} - \widehat{Y}\|_2^2 = 4\big[d^2 - \big\|Y - \tfrac{1}{2}(\overline{Y} + \widehat{Y})\big\|_2^2\big] \leq 4(d^2 - d^2) = 0.$$

Hence $\mathbb{E}((\overline{Y} - \widehat{Y})^2) = 0$ and so $\mathbb{P}(\widehat{Y} = \overline{Y}) = 1$. ■

(12) Example. Let Y have mean μ and variance σ^2. With no information about Y, it is appropriate to ask for the real number h which minimizes $\|Y - h\|_2$. Now $\|Y - h\|_2^2 = \mathbb{E}((Y - h)^2) = \sigma^2 + (\mu - h)^2$, so that μ is the best predictor of Y. The set H of possible estimators is the real line $\mathbb{R}$. ●

† $\|Z\|_2^2$ denotes $\{\|Z\|_2\}^2$.

(13) Example. Let $X_1, X_2, \ldots$ be uncorrelated random variables with zero means and unit variances. It is desired to find the best predictor of Y amongst the class of linear combinations of the X_i. Clearly

$$\mathbb{E}\left(\left(Y - \sum_i a_i X_i\right)^2\right) = \mathbb{E}(Y^2) - 2\sum_i a_i \mathbb{E}(X_i Y) + \sum_i a_i^2$$
$$= \mathbb{E}(Y^2) + \sum_i \left[a_i - \mathbb{E}(X_i Y)\right]^2 - \sum_i \mathbb{E}(X_i Y)^2.$$

This is a minimum when $a_i = \mathbb{E}(X_i Y)$ for all i, so that $\widehat{Y} = \sum_i X_i \mathbb{E}(X_i Y)$. (*Exercise*: Prove that $\mathbb{E}(\widehat{Y}^2) < \infty$.) This is seen best in the following light. Thinking of the X_i as orthogonal (that is, uncorrelated) unit vectors in the space H of linear combinations of the X_i, we have found that $\widehat{Y}$ is the weighted average of the X_i, weighted in proportion to the magnitudes of their 'projections' onto Y. The geometry of this example is relevant to Theorem (14). ●

(14) Projection theorem. *Let H be a closed linear space (with respect to $\|\cdot\|_2$) of random variables, and let Y satisfy $\mathbb{E}(Y^2) < \infty$. Let $M \in H$. The following two statements are equivalent*:

(15) $$\mathbb{E}\big((Y - M)Z\big) = 0 \quad \textit{for all } Z \in H,$$

(16) $$\|Y - M\|_2 \leq \|Y - Z\|_2 \quad \textit{for all } Z \in H.$$

Here is the geometrical intuition. Let $L_2(\Omega, \mathcal{F}, \mathbb{P})$ be the set of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ having finite second moment. Now H is a linear subspace of $L_2(\Omega, \mathcal{F}, \mathbb{P})$; think of H as a hyperplane in a vector space of very large dimension. If $Y \notin H$ then the shortest route from Y to H is along the perpendicular from Y onto H. Writing $\widehat{Y}$ for the foot of this perpendicular, we have that $Y - \widehat{Y}$ is perpendicular to any vector in the hyperplane H. Translating this geometrical remark back into the language of random variables, we conclude that $\langle Y - \widehat{Y}, Z\rangle = 0$ for all $Z \in H$, where $\langle U, V\rangle$ is the scalar product in $L_2(\Omega, \mathcal{F}, \mathbb{P})$ defined by $\langle U, V\rangle = \mathbb{E}(UV)$. These remarks do not of course constitute a proof of the theorem.

Proof. Suppose first that $M \in H$ satisfies (15). Then, for $M' \in H$,

$$\mathbb{E}\big((Y - M')^2\big) = \mathbb{E}\big((Y - M + M - M')^2\big)$$
$$= \mathbb{E}\big((Y - M)^2\big) + \mathbb{E}\big((M - M')^2\big)$$

by (15), since $M - M' \in H$; therefore $\|Y - M\|_2 \leq \|Y - M'\|_2$ for all $M' \in H$.

Conversely, suppose that M satisfies (16), but that there exists $Z \in H$ such that

$$\mathbb{E}\big((Y - M)Z\big) = d > 0.$$

We may assume without loss of generality that $\mathbb{E}(Z^2) = 1$; otherwise replace Z by $Z/\sqrt{\mathbb{E}(Z^2)}$, noting that $\mathbb{E}(Z^2) \neq 0$ since $\mathbb{P}(Z = 0) \neq 1$. Writing $M' = M + dZ$, we have that

$$\mathbb{E}\big((Y - M')^2\big) = \mathbb{E}\big((Y - M + M - M')^2\big)$$
$$= \mathbb{E}\big((Y - M)^2\big) - 2d\mathbb{E}\big((Y - M)Z\big) + d^2\mathbb{E}(Z^2)$$
$$= \mathbb{E}\big((Y - M)^2\big) - d^2,$$

in contradiction of the minimality of $\mathbb{E}((Y - M)^2)$. ■

It is only a tiny step from the projection theorem (14) to the observation, well known to statisticians, that the best predictor of Y given X is just the conditional expectation $\mathbb{E}(Y \mid X)$. This fact, easily proved directly (*exercise*), follows immediately from the projection theorem.

(**17**) **Theorem.** *Let X and Y be random variables, and suppose that $\mathbb{E}(Y^2) < \infty$. The best predictor of Y given X is the conditional expectation $\mathbb{E}(Y \mid X)$.*

Proof. Let H be the closed linear space of functions of X having finite second moment. Define $\psi(x) = \mathbb{E}(Y \mid X = x)$. Certainly $\psi(X)$ belongs to H, since

$$\mathbb{E}(\psi(X)^2) = \mathbb{E}\big(\mathbb{E}(Y \mid X)^2\big) \leq \mathbb{E}\big(\mathbb{E}(Y^2 \mid X)\big) = \mathbb{E}(Y^2),$$

where we have used the Cauchy–Schwarz inequality. On the other hand, for $Z = h(X) \in H$,

$$\begin{aligned}\mathbb{E}\big([Y - \psi(X)]Z\big) &= \mathbb{E}(Yh(X)) - \mathbb{E}\big(\mathbb{E}(Y \mid X)h(X)\big) \\ &= \mathbb{E}(Yh(X)) - \mathbb{E}\big(\mathbb{E}(Yh(X) \mid X)\big) \\ &= \mathbb{E}(Yh(X)) - \mathbb{E}\big(Yh(X)\big) = 0,\end{aligned}$$

using the elementary fact that $\mathbb{E}(Yh(X) \mid X) = h(X)\mathbb{E}(Y \mid X)$. Applying the projection theorem, we find that $M = \psi(X)$ $(= \mathbb{E}(Y \mid X))$ minimizes $\|Y - M\|_2$ for $M \in H$, which is the claim of the theorem. ■

Here is an important step. We may take the conclusion of (17) as a *definition* of the conditional expectation $\mathbb{E}(Y \mid X)$: if $\mathbb{E}(Y^2) < \infty$, the *conditional expectation* $\mathbb{E}(Y \mid X)$ of Y given X is defined to be the best predictor of Y given X.

There are two major advantages of defining conditional expectation in this way. First, it is a definition which is valid for all pairs X, Y such that $\mathbb{E}(Y^2) < \infty$, regardless of their types (discrete, continuous, and so on). Secondly, it provides a route to a much more general notion of conditional expectation which turns out to be particularly relevant to the martingale theory of Chapter 12.

(**18**) **Example.** Let $X = \{X_i : i \in I\}$ be a family of random variables, and let H be the space of all functions of the X_i with finite second moments. If $\mathbb{E}(Y^2) < \infty$, the *conditional expectation* $\mathbb{E}(Y \mid X_i,\ i \in I)$ of Y given the X_i is defined to be the function $M = \psi(X) \in H$ which minimizes the mean squared error $\|Y - M\|_2$ over all M in H. Note that $\psi(X)$ satisfies

$$\mathbb{E}\big([Y - \psi(X)]Z\big) = 0 \quad \text{for all } Z \in H, \tag{19}$$

and $\psi(X)$ is unique in the sense that $\mathbb{P}(\psi(X) = N) = 1$ if $\|Y - \psi(X)\|_2 = \|Y - N\|_2$ for any $N \in H$. We note here that, strictly speaking, conditional expectations are not actually *unique*; this causes no difficulty, and we shall therefore continue to speak in terms of *the* conditional expectation. ●

We move on to an important generalization of the idea of conditional expectation, involving 'conditioning on a σ-field'. Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ having finite second moment, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let H be the space of random variables which are $\mathcal{G}$-measurable and have finite second moment. That is to say, H contains those random variables Z such that $\mathbb{E}(Z^2) < \infty$ and $\{Z \leq z\} \in \mathcal{G}$ for all $z \in \mathbb{R}$. It is not difficult to see that H is a closed linear space with respect to $\|\cdot\|_2$. We have from (9) that there exists an element

M of H such that $\|Y - M\|_2 \leq \|Y - Z\|_2$ for all $Z \in H$, and furthermore M is unique (in the usual way) with this property. We call M the 'conditional expectation of Y given the σ-field $\mathcal{G}$', written $\mathbb{E}(Y \mid \mathcal{G})$.

This is a more general definition of conditional expectation than that obtained by conditioning on a family of random variables (as in the previous example). To see this, take $\mathcal{G}$ to be the smallest σ-field with respect to which every member of the family $X = \{X_i : i \in I\}$ is measurable. It is clear that $\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(Y \mid X_i,\ i \in I)$, in the sense that they are equal with probability 1.

We arrive at the following definition by use of the projection theorem (14).

(20) Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let Y be a random variable satisfying $\mathbb{E}(Y^2) < \infty$. If $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$, the **conditional expectation** $\mathbb{E}(Y \mid \mathcal{G})$ is a $\mathcal{G}$-measurable random variable satisfying

$$\mathbb{E}\big(\big[Y - \mathbb{E}(Y \mid \mathcal{G})\big]Z\big) = 0 \quad \text{for all } Z \in H, \tag{21}$$

where H is the collection of all $\mathcal{G}$-measurable random variables with finite second moment.

There are certain members of H with particularly simple form, being the indicator functions of events in $\mathcal{G}$. It may be shown without great difficulty that condition (21) may be replaced by

$$\mathbb{E}\big(\big[Y - \mathbb{E}(Y \mid \mathcal{G})\big]I_G\big) = 0 \quad \text{for all } G \in \mathcal{G}. \tag{22}$$

Setting $G = \Omega$, we deduce the important fact that

$$\mathbb{E}\big(\mathbb{E}(Y \mid \mathcal{G})\big) = \mathbb{E}(Y). \tag{23}$$

(24) Example. Doob's martingale. (Though some ascribe this to Lévy.) Let Y have finite second moment, and let $X_1, X_2, \ldots$ be a sequence of random variables. Define

$$Y_n = \mathbb{E}(Y \mid X_1, X_2, \ldots, X_n).$$

Then $\{Y_n\}$ is a martingale with respect to $\{X_n\}$. To show this it is necessary to prove that $\mathbb{E}|Y_n| < \infty$ and $\mathbb{E}(Y_{n+1} \mid X_1, X_2, \ldots, X_n) = Y_n$. Certainly $\mathbb{E}|Y_n| < \infty$, since $\mathbb{E}(Y_n^2) < \infty$. For the other part, let H_n be the space of functions of $X_1, X_2, \ldots, X_n$ having finite second moment. We have by (19) that, for $Z \in H_n$,

$$\begin{aligned} 0 &= \mathbb{E}\big((Y - Y_n)Z\big) = \mathbb{E}\big((Y - Y_{n+1} + Y_{n+1} - Y_n)Z\big) \\ &= \mathbb{E}\big((Y_{n+1} - Y_n)Z\big) \quad \text{since} \quad Z \in H_n \subseteq H_{n+1}. \end{aligned}$$

Therefore $Y_n = \mathbb{E}(Y_{n+1} \mid X_1, X_2, \ldots, X_n)$.

Here is a more general formulation. Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}(Y^2) < \infty$, and let $\{\mathcal{G}_n : n \geq 1\}$ be a sequence of σ-fields contained in $\mathcal{F}$ and satisfying $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$ for all n. Such a sequence $\{\mathcal{G}_n\}$ is called a *filtration*; in the context of the previous paragraph we might take $\mathcal{G}_n$ to be the smallest σ-field with respect to which $X_1, X_2, \ldots, X_n$ are each measurable. We define $Y_n = \mathbb{E}(Y \mid \mathcal{G}_n)$. As before $\{Y_n\}$ satisfies $\mathbb{E}|Y_n| < \infty$ and $\mathbb{E}(Y_{n+1} \mid \mathcal{G}_n) = Y_n$; such a sequence is called a 'martingale with respect to the filtration $\{\mathcal{G}_n\}$'. ●

This new type of conditional expectation has many useful properties. We single out one of these, namely the *pull-through property*, thus named since it involves a random variable being pulled through a paranthesis.

(25) Theorem. *Let Y have finite second moment and let $\mathcal{G}$ be a sub-σ-field of the σ-field $\mathcal{F}$. Then $\mathbb{E}(XY \mid \mathcal{G}) = X\mathbb{E}(Y \mid \mathcal{G})$ for all $\mathcal{G}$-measurable random variables X with finite second moments.*

Proof. Let X be $\mathcal{G}$-measurable with finite second moment. Clearly

$$Z = \mathbb{E}(XY \mid \mathcal{G}) - X\mathbb{E}(Y \mid \mathcal{G})$$

is $\mathcal{G}$-measurable and satisfies

$$Z = X\big[Y - \mathbb{E}(Y \mid \mathcal{G})\big] - \big[XY - \mathbb{E}(XY \mid \mathcal{G})\big]$$

so that, for $G \in \mathcal{G}$,

$$\mathbb{E}(ZI_G) = \mathbb{E}\big(\big[Y - \mathbb{E}(Y \mid \mathcal{G})\big]XI_G\big) - \mathbb{E}\big(\big[XY - \mathbb{E}(XY \mid \mathcal{G})\big]I_G\big) = 0,$$

the first term being zero by the fact that XI_G is $\mathcal{G}$-measurable with finite second moment, and the second by the definition of $\mathbb{E}(XY \mid \mathcal{G})$. Any $\mathcal{G}$-measurable random variable Z satisfying $\mathbb{E}(ZI_G) = 0$ for all $G \in \mathcal{G}$ is such that $\mathbb{P}(Z = 0) = 1$ (just set $G_1 = \{Z > 0\}$, $G_2 = \{Z < 0\}$ in turn), and the result follows. ∎

In all our calculations so far, we have used the norm $\|\cdot\|_2$, leading to a definition of $\mathbb{E}(Y \mid \mathcal{G})$ for random variables Y with $\mathbb{E}(Y^2) < \infty$. This condition of finite second moment is of course too strong in general, and needs to be replaced by the natural weaker condition that $\mathbb{E}|Y| < \infty$. One way of doing this would be to rework the previous arguments using instead the norm $\|\cdot\|_1$. An easier route is to use the technique of 'truncation' as in the following proof.

(26) Theorem. *Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}|Y| < \infty$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. There exists a random variable Z such that*:

(a) *Z is $\mathcal{G}$-measurable,*

(b) $\mathbb{E}|Z| < \infty$,

(c) $\mathbb{E}((Y - Z)I_G) = 0$ *for all* $G \in \mathcal{G}$.

Z is unique in the sense that, for any Z' satisfying (a), (b), *and* (c), *we have that* $\mathbb{P}(Z = Z') = 1$.

The random variable Z in the theorem is called the 'conditional expectation of Y given $\mathcal{G}$', and is written $\mathbb{E}(Y \mid \mathcal{G})$. It is an *exercise* to prove that

$$\mathbb{E}(XY \mid \mathcal{G}) = X\mathbb{E}(Y \mid \mathcal{G}) \tag{27}$$

for all $\mathcal{G}$-measurable X, whenever both sides exist, and also that this definition coincides (almost surely) with the previous one when Y has finite second moment. A meaningful value can be assigned to $\mathbb{E}(Y \mid \mathcal{G})$ under the weaker assumption on Y that either $\mathbb{E}(Y^+) < \infty$ or $\mathbb{E}(Y^-) < \infty$.

Proof. Suppose first that $Y \geq 0$ and $\mathbb{E}|Y| < \infty$. Let $Y_n = \min\{Y, n\}$, so that $Y_n \uparrow Y$ as $n \to \infty$. Certainly $\mathbb{E}(Y_n^2) < \infty$, and hence we may use (20) to find the conditional expectation $\mathbb{E}(Y_n \mid \mathcal{G})$, a $\mathcal{G}$-measurable random variable satisfying

$$\mathbb{E}\big([Y_n - \mathbb{E}(Y_n \mid \mathcal{G})]I_G\big) = 0 \quad \text{for all } G \in \mathcal{G}. \tag{28}$$

Now $Y_n \leq Y_{n+1}$, and so we may take $\mathbb{E}(Y_n \mid \mathcal{G}) \leq \mathbb{E}(Y_{n+1} \mid \mathcal{G})$; see Exercise (7.9.4iii). Hence $\lim_{n\to\infty} \mathbb{E}(Y_n \mid \mathcal{G})$ exists, and we write $\mathbb{E}(Y \mid \mathcal{G})$ for this limit, a $\mathcal{G}$-measurable random variable. By monotone convergence (5.6.12) and (23), $\mathbb{E}(Y_n I_G) \uparrow \mathbb{E}(Y I_G)$, and

$$\mathbb{E}\big[\mathbb{E}(Y_n \mid \mathcal{G})I_G\big] \uparrow \mathbb{E}\big[\mathbb{E}(Y \mid \mathcal{G})I_G\big] = \mathbb{E}\big[\mathbb{E}(Y I_G \mid \mathcal{G})\big] = \mathbb{E}(Y I_G),$$

so that, by (28), $\mathbb{E}([Y - \mathbb{E}(Y \mid \mathcal{G})]I_G) = 0$ for all $G \in \mathcal{G}$.

Next we lift in the usual way the restriction that Y be non-negative. We express Y as $Y = Y^+ - Y^-$ where $Y^+ = \max\{Y, 0\}$ and $Y^- = -\min\{Y, 0\}$ are non-negative; we define $\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(Y^+ \mid \mathcal{G}) - \mathbb{E}(Y^- \mid \mathcal{G})$. It is easy to check that $\mathbb{E}(Y \mid \mathcal{G})$ satisfies (a), (b), and (c). To see the uniqueness, suppose that there exist two $\mathcal{G}$-measurable random variables Z_1 and Z_2 satisfying (c). Then $\mathbb{E}((Z_1 - Z_2)I_G) = \mathbb{E}((Y - Y)I_G) = 0$ for all $G \in \mathcal{G}$. Setting $G = \{Z_1 > Z_2\}$ and $G = \{Z_1 < Z_2\}$ in turn, we find that $\mathbb{P}(Z_1 = Z_2) = 1$ as required. ■

Having defined $\mathbb{E}(Y \mid \mathcal{G})$, we can of course define conditional probabilities also: if $A \in \mathcal{F}$, we define $\mathbb{P}(A \mid \mathcal{G}) = \mathbb{E}(I_A \mid \mathcal{G})$. It may be checked that $\mathbb{P}(\varnothing \mid \mathcal{G}) = 0$, $\mathbb{P}(\Omega \mid \mathcal{G}) = 1$ a.s., and $\mathbb{P}(\bigcup_i A_i \mid \mathcal{G}) = \sum_i \mathbb{P}(A_i \mid \mathcal{G})$ a.s. for any sequence $\{A_i : i \geq 1\}$ of disjoint events in $\mathcal{F}$.

It looks as though there should be a way of defining $\mathbb{P}(\cdot \mid \mathcal{G})$ so that it is a probability measure on $(\Omega, \mathcal{F})$. This turns out to be impossible in general, but the details are beyond the scope of this book.

Exercises for Section 7.9

1. Let Y be uniformly distributed on $[-1, 1]$ and let $X = Y^2$.
(a) Find the best predictor of X given Y, and of Y given X.
(b) Find the best linear predictor of X given Y, and of Y given X.

2. Let the pair (X, Y) have a general bivariate normal distribution. Find $\mathbb{E}(Y \mid X)$.

3. Let $X_1, X_2, \ldots, X_n$ be random variables with zero means and covariance matrix $\mathbf{V} = (v_{ij})$, and let Y have finite second moment. Find the linear function h of the X_i which minimizes the mean squared error $\mathbb{E}\{(Y - h(X_1, \ldots, X_n))^2\}$.

4. Verify the following properties of conditional expectation. You may assume that the relevant expectations exist.
(i) $\mathbb{E}\{\mathbb{E}(Y \mid \mathcal{G})\} = \mathbb{E}(Y)$.
(ii) $\mathbb{E}(\alpha Y + \beta Z \mid \mathcal{G}) = \alpha\mathbb{E}(Y \mid \mathcal{G}) + \beta\mathbb{E}(Z \mid \mathcal{G})$ for $\alpha, \beta \in \mathbb{R}$.
(iii) $\mathbb{E}(Y \mid \mathcal{G}) \geq 0$ if $Y \geq 0$.
(iv) $\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}\{\mathbb{E}(Y \mid \mathcal{H}) \mid \mathcal{G}\}$ if $\mathcal{G} \subseteq \mathcal{H}$.
(v) $\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(Y)$ if Y is independent of I_G for every $G \in \mathcal{G}$.
(vi) **Jensen's inequality.** $g\{\mathbb{E}(Y \mid \mathcal{G})\} \leq \mathbb{E}\{g(Y) \mid \mathcal{G}\}$ for all convex functions g.
(vii) If $Y_n \xrightarrow{\text{a.s.}} Y$ and $|Y_n| \leq Z$ a.s. where $\mathbb{E}(Z) < \infty$, then $\mathbb{E}(Y_n \mid \mathcal{G}) \xrightarrow{\text{a.s.}} \mathbb{E}(Y \mid \mathcal{G})$.
(Statements (ii)–(vi) are of course to be interpreted 'almost surely'.)

5. Let X and Y have joint mass function $f(x, y) = \{x(x + 1)\}^{-1}$ for $x = y = 1, 2, \ldots$. Show that $\mathbb{E}(Y \mid X) < \infty$ while $\mathbb{E}(Y) = \infty$.

6. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let H be the space of $\mathcal{G}$-measurable random variables with finite second moment.

(a) Show that H is closed with respect to the norm $\|\cdot\|_2$.

(b) Let Y be a random variable satisfying $\mathbb{E}(Y^2) < \infty$, and show the equivalence of the following two statements for any $M \in H$:

(i) $\mathbb{E}\{(Y - M)Z\} = 0$ for all $Z \in H$,

(ii) $\mathbb{E}\{(Y - M)I_G\} = 0$ for all $G \in \mathcal{G}$.

7.10 Uniform integrability

Suppose that we are presented with a sequence $\{X_n : n \geq 1\}$ of random variables, and we are able to prove that $X_n \xrightarrow{\mathrm{P}} X$. Convergence in probability tells us little about the behaviour of $\mathbb{E}(X_n)$, as the trite example

$$Y_n = \begin{cases} n & \text{with probability } n^{-1}, \\ 0 & \text{otherwise,} \end{cases}$$

shows; in this special case, $Y_n \xrightarrow{\mathrm{P}} 0$ but $\mathbb{E}(Y_n) = 1$ for all n. Should we wish to prove that $\mathbb{E}(X_n) \to \mathbb{E}(X)$, or further that $X_n \xrightarrow{1} X$ (which is to say that $\mathbb{E}|X_n - X| \to 0$), then an additional condition is required.

We encountered in an earlier Exercise (7.2.2) an argument of the kind required. If $X_n \xrightarrow{\mathrm{P}} X$ and $|X_n| \leq Y$ for some Y such that $\mathbb{E}|Y| < \infty$, then $X_n \xrightarrow{1} X$. This extra condition, that $\{X_n\}$ be dominated *uniformly*, is often too strong an assumption in cases of interest. A weaker condition is provided by the following definition. As usual I_A denotes the indicator function of the event A.

(1) Definition. A sequence $X_1, X_2, \ldots$ of random variables is said to be **uniformly integrable** if

$$\sup_n \mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) \to 0 \quad \text{as} \quad a \to \infty. \tag{2}$$

Let us investigate this condition briefly. A random variable Y is called 'integrable' if $\mathbb{E}|Y| < \infty$, which is to say that

$$\mathbb{E}(|Y| I_{\{|Y| \geq a\}}) = \int_{|y| \geq a} |y| \, dF_Y(y)$$

tends to 0 as $a \to \infty$ (see Exercise (5.6.5)). Therefore, a family $\{X_n : n \geq 1\}$ is 'integrable' if

$$\mathbb{E}(|X_n| I_{\{X_n \geq a\}}) \to 0 \quad \text{as} \quad a \to \infty$$

for all n, and is 'uniformly integrable' if the convergence is uniform in n. Roughly speaking, the condition of integrability restricts the amount of probability in the tails of the distribution, and uniform integrability restricts such quantities *uniformly* over the family of random variables in question.

The principal use of uniform integrability is demonstrated by the following theorem.

(3) Theorem. *Suppose that* $X_1, X_2, \ldots$ *is a sequence of random variables satisfying* $X_n \xrightarrow{\mathrm{P}} X$. *The following three statements are equivalent to one another.*

(a) *The family* $\{X_n : n \geq 1\}$ *is uniformly integrable.*

(b) $\mathbb{E}|X_n| < \infty$ *for all* n, $\mathbb{E}|X| < \infty$, *and* $X_n \xrightarrow{1} X$.

(c) $\mathbb{E}|X_n| < \infty$ *for all* n, *and* $\mathbb{E}|X_n| \to \mathbb{E}|X| < \infty$.

In advance of proving this, we note some sufficient conditions for uniform integrability.

(4) Example. Suppose $|X_n| \leq Y$ for all n, where $\mathbb{E}|Y| < \infty$. Then

$$|X_n| I_{\{|X_n| \geq a\}} \leq |Y| I_{\{|Y| \geq a\}},$$

so that

$$\sup_n \mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) \leq \mathbb{E}(|Y| I_{\{|Y| \geq a\}})$$

which tends to zero as $a \to \infty$, since $\mathbb{E}|Y| < \infty$. ●

(5) Example. Suppose that there exist $\delta > 0$ and $K < \infty$ such that $\mathbb{E}(|X_n|^{1+\delta}) \leq K$ for all n. Then

$$\begin{aligned} \mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) &\leq \frac{1}{a^\delta} \mathbb{E}(|X_n|^{1+\delta} I_{\{|X_n| \geq a\}}) \\ &\leq \frac{1}{a^\delta} \mathbb{E}(|X_n|^{1+\delta}) \leq \frac{K}{a^\delta} \to 0 \end{aligned}$$

as $a \to \infty$, so that the family is uniformly integrable. ●

Turning to the proof of Theorem (3), we note first a preliminary lemma which is of value in its own right.

(6) Lemma. *A family* $\{X_n : n \geq 1\}$ *is uniformly integrable if and only if both of the following hold*:

(a) $\sup_n \mathbb{E}|X_n| < \infty$,

(b) *for all* $\epsilon > 0$, *there exists* $\delta > 0$ *such that, for all* n, $\mathbb{E}(|X_n| I_A) < \epsilon$ *for any event* A *such that* $\mathbb{P}(A) < \delta$.

The equivalent statement for a single random variable X is the assertion that $\mathbb{E}|X| < \infty$ if and only if

$$\sup_{A : \mathbb{P}(A) < \delta} \mathbb{E}(|X| I_A) \to 0 \quad \text{as} \quad \delta \to 0; \tag{7}$$

see Exercise (5.6.5).

Proof of (6). Suppose first that $\{X_n\}$ is uniformly integrable. For any $a > 0$,

$$\mathbb{E}|X_n| = \mathbb{E}(|X_n| I_{\{X_n < a\}}) + \mathbb{E}(|X_n| I_{\{X_n \geq a\}}),$$

and therefore

$$\sup_n \mathbb{E}|X_n| \le a + \sup_n \mathbb{E}(|X_n| I_{\{X_n \ge a\}}).$$

We use uniform integrability to find that $\sup_n \mathbb{E}|X_n| < \infty$. Next,

$$\mathbb{E}(|X_n| I_A) = \mathbb{E}(|X_n| I_{A \cap B_n(a)}) + \mathbb{E}(|X_n| I_{A \cap B_n(a)^c}) \tag{8}$$

where $B_n(a) = \{|X_n| \ge a\}$. Now

$$\mathbb{E}(|X_n| I_{A \cap B_n(a)}) \le \mathbb{E}(|X_n| I_{B_n(a)})$$

and

$$\mathbb{E}(|X_n| I_{A \cap B_n(a)^c}) \le a\mathbb{E}(I_A) = a\mathbb{P}(A).$$

Let $\epsilon > 0$ and pick a such that $\mathbb{E}(|X_n| I_{B_n(a)}) < \frac{1}{2}\epsilon$ for all n. We have from (8) that $\mathbb{E}(|X_n| I_A) \le \frac{1}{2}\epsilon + a\mathbb{P}(A)$, which is smaller than ϵ whenever $\mathbb{P}(A) < \epsilon/(2a)$.

Secondly, suppose that (a) and (b) hold; let $\epsilon > 0$ and pick δ according to (b). We have that

$$\mathbb{E}|X_n| \ge \mathbb{E}(|X_n| I_{B_n(a)}) \ge a\mathbb{P}(B_n(a))$$

(this is Markov's inequality) so that

$$\sup_n \mathbb{P}(B_n(a)) \le \frac{1}{a} \sup_n \mathbb{E}|X_n| < \infty.$$

Pick a such that $a^{-1} \sup_n \mathbb{E}|X_n| < \delta$, implying that $\mathbb{P}(B_n(a)) < \delta$ for all n. It follows from (b) that $\mathbb{E}(|X_n| I_{B_n(a)}) < \epsilon$ for all n, and hence $\{X_n\}$ is uniformly integrable. ■

Proof of Theorem (3). The main part is the statement that (a) implies (b), and we prove this first. Suppose that the family is uniformly integrable. Certainly each member is integrable, so that $\mathbb{E}|X_n| < \infty$ for all n. Since $X_n \xrightarrow{\mathrm{P}} X$, there exists a subsequence $\{X_{n_k} : k \ge 1\}$ such that $X_{n_k} \xrightarrow{\text{a.s.}} X$ (see Theorem (7.2.13)). By Fatou's lemma (5.6.13),

$$\mathbb{E}|X| = \mathbb{E}\big(\liminf_{k\to\infty} |X_{n_k}|\big) \le \liminf_{k\to\infty} \mathbb{E}|X_{n_k}| \le \sup_n \mathbb{E}|X_n|, \tag{9}$$

which is finite as a consequence of Lemma (6).

To prove convergence in mean, we write, for $\epsilon > 0$,

$$\begin{aligned} \mathbb{E}|X_n - X| &= \mathbb{E}\Big(|X_n - X| I_{\{|X_n - X| < \epsilon\}} + |X_n - X| I_{\{|X_n - X| \ge \epsilon\}}\Big) \\ &\le \epsilon + \mathbb{E}(|X_n| I_{A_n}) + \mathbb{E}(|X| I_{A_n}) \end{aligned} \tag{10}$$

where $A_n = \{|X_n - X| > \epsilon\}$. Now $\mathbb{P}(A_n) \to 0$ in the limit as $n \to \infty$, and hence $\mathbb{E}(|X_n| I_{A_n}) \to 0$ as $n \to \infty$, by Lemma (6). Similarly $\mathbb{E}(|X| I_{A_n}) \to 0$ as $n \to \infty$, by (7), so that $\limsup_{n\to\infty} \mathbb{E}|X_n - X| \le \epsilon$. Let $\epsilon \downarrow 0$ to obtain that $X_n \xrightarrow{1} X$.

That (b) implies (c) is immediate from the observation that

$$\big|\mathbb{E}|X_n| - \mathbb{E}|X|\big| \le \mathbb{E}|X_n - X|,$$

and it remains to prove that (c) implies (a). Suppose then that (c) holds. Clearly

$$\mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) = \mathbb{E}|X_n| - \mathbb{E}(u(X_n)) \tag{11}$$

where $u(x) = |x| I_{(-a,a)}(x)$. Now u is a continuous bounded function on $(-a, a)$ and $X_n \xrightarrow{\mathrm{D}} X$; hence

$$\mathbb{E}(u(X_n)) \to \mathbb{E}(u(X)) = \mathbb{E}(|X| I_{\{|X|<a\}})$$

if a and $-a$ are points of continuity of the distribution function F_X of X (see Theorem (7.2.19) and the comment thereafter). The function F_X is monotone, and therefore the set Δ of discontinuities of F_X is at most countable. It follows from (11) that

$$\mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) \to \mathbb{E}|X| - \mathbb{E}(|X| I_{\{|X|<a\}}) = \mathbb{E}(|X| I_{\{|X| \geq a\}}) \tag{12}$$

if $a \notin \Delta$. For any $\epsilon > 0$, on the one hand there exists $b \notin \Delta$ such that $\mathbb{E}(|X| I_{\{|X| \geq b\}}) < \epsilon$; with this choice of b, there exists by (12) an integer N such that $\mathbb{E}(|X_n| I_{\{|X_n| \geq b\}}) < 2\epsilon$ for all $n \geq N$. On the other hand, there exists c such that $\mathbb{E}(|X_k| I_{\{|X_k| \geq c\}}) < 2\epsilon$ for all $k < N$, since only finitely many terms are involved. If $a > \max\{b, c\}$, we have that $\mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) < 2\epsilon$ for all n, and we have proved that $\{X_n\}$ is uniformly integrable. ■

The concept of uniform integrability will be of particular value when we return in Chapter 12 to the theory of martingales. The following example may be seen as an illustration of this.

(13) Example. Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}|Y| < \infty$, and let $\{\mathcal{G}_n : n \geq 1\}$ be a filtration, which is to say that $\mathcal{G}_n$ is a sub-σ-field of $\mathcal{F}$, and furthermore $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$ for all n. Let $X_n = \mathbb{E}(Y \mid \mathcal{G}_n)$. The sequence $\{X_n : n \geq 1\}$ is uniformly integrable, as may be seen in the following way.

It is a consequence of Jensen's inequality, Exercise (7.9.4vi), that

$$|X_n| = \left|\mathbb{E}(Y \mid \mathcal{G}_n)\right| \leq \mathbb{E}\left(|Y| \,\middle|\, \mathcal{G}_n\right)$$

almost surely, so that $\mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) \leq \mathbb{E}(Z_n I_{\{Z_n \geq a\}})$ where $Z_n = \mathbb{E}\left(|Y| \,\middle|\, \mathcal{G}_n\right)$. By the definition of conditional expectation, $\mathbb{E}\left\{(|Y| - Z_n) I_{\{Z_n \geq a\}}\right\} = 0$, so that

$$\mathbb{E}(|X_n| I_{\{|X_n| \geq a\}}) \leq \mathbb{E}(|Y| I_{\{Z_n \geq a\}}). \tag{14}$$

We now repeat an argument used before. By Markov's inequality,

$$\mathbb{P}(Z_n \geq a) \leq a^{-1}\mathbb{E}(Z_n) = a^{-1}\mathbb{E}|Y|,$$

and therefore $\mathbb{P}(Z_n \geq a) \to 0$ as $a \to \infty$, uniformly in n. Using (7), we deduce that $\mathbb{E}(|Y| I_{\{Z_n \geq a\}}) \to 0$ as $a \to \infty$, uniformly in n, implying that the sequence $\{X_n\}$ is uniformly integrable. ●

We finish this section with an application.

(15) Example. Convergence of moments. Suppose that $X_1, X_2, \ldots$ is a sequence satisfying $X_n \xrightarrow{\mathrm{D}} X$, and furthermore $\sup_n \mathbb{E}(|X_n|^\alpha) < \infty$ for some $\alpha > 1$. It follows that

$$\mathbb{E}(X_n^\beta) \to \mathbb{E}(X^\beta) \tag{16}$$

for any integer β satisfying $1 \leq \beta < \alpha$. This may be proved either directly or via Theorem (3). First, if β is an integer satisfying $1 \leq \beta < \alpha$, then $\{X_n^\beta : n \geq 1\}$ is uniformly integrable by (5), and furthermore $X_n^\beta \xrightarrow{\mathrm{D}} X^\beta$ (easy *exercise*, or use Theorem (7.2.18)). If it were the case that $X_n^\beta \xrightarrow{\mathrm{P}} X^\beta$ then Theorem (3) would imply the result. In any case, by the Skorokhod representation theorem (7.2.14), there exist random variables $Y, Y_1, Y_2, \ldots$ having the same distributions as $X, X_1, X_2, \ldots$ such that $Y_n^\beta \xrightarrow{\mathrm{P}} Y^\beta$. Thus $\mathbb{E}(Y_n^\beta) \to \mathbb{E}(Y^\beta)$ by Theorem (3). However, $\mathbb{E}(Y_n^\beta) = \mathbb{E}(X_n^\beta)$ and $\mathbb{E}(Y^\beta) = \mathbb{E}(X^\beta)$, and the proof is complete. ●

Exercises for Section 7.10

1. Show that the sum $\{X_n + Y_n\}$ of two uniformly integrable sequences $\{X_n\}$ and $\{Y_n\}$ gives a uniformly integrable sequence.

2. (a) Suppose that $X_n \xrightarrow{r} X$ where $r \geq 1$. Show that $\{|X_n|^r : n \geq 1\}$ is uniformly integrable, and deduce that $\mathbb{E}(X_n^r) \to \mathbb{E}(X^r)$ if r is an integer.

(b) Conversely, suppose that $\{|X_n|^r : n \geq 1\}$ is uniformly integrable where $r \geq 1$, and show that $X_n \xrightarrow{r} X$ if $X_n \xrightarrow{\mathrm{P}} X$.

3. Let $g : [0, \infty) \to [0, \infty)$ be an increasing function satisfying $g(x)/x \to \infty$ as $x \to \infty$. Show that the sequence $\{X_n : n \geq 1\}$ is uniformly integrable if $\sup_n \mathbb{E}\{g(|X_n|)\} < \infty$.

4. Let $\{Z_n : n \geq 0\}$ be the generation sizes of a branching process with $Z_0 = 1$, $\mathbb{E}(Z_1) = 1$, $\mathrm{var}(Z_1) \neq 0$. Show that $\{Z_n : n \geq 0\}$ is not uniformly integrable.

5. Pratt's lemma. Suppose that $X_n \leq Y_n \leq Z_n$ where $X_n \xrightarrow{\mathrm{P}} X$, $Y_n \xrightarrow{\mathrm{P}} Y$, and $Z_n \xrightarrow{\mathrm{P}} Z$. If $\mathbb{E}(X_n) \to \mathbb{E}(X)$ and $\mathbb{E}(Z_n) \to \mathbb{E}(Z)$, show that $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$.

6. Let $\{X_n : n \geq 1\}$ be a sequence of variables satisfying $\mathbb{E}(\sup_n |X_n|) < \infty$. Show that $\{X_n\}$ is uniformly integrable.

7.11 Problems

1. Let X_n have density function

$$f_n(x) = \frac{n}{\pi(1 + n^2x^2)}, \qquad n \geq 1.$$

With respect to which modes of convergence does X_n converge as $n \to \infty$?

2. (i) Suppose that $X_n \xrightarrow{\text{a.s.}} X$ and $Y_n \xrightarrow{\text{a.s.}} Y$, and show that $X_n + Y_n \xrightarrow{\text{a.s.}} X + Y$. Show that the corresponding result holds for convergence in rth mean and in probability, but not in distribution.

(ii) Show that if $X_n \xrightarrow{\text{a.s.}} X$ and $Y_n \xrightarrow{\text{a.s.}} Y$ then $X_nY_n \xrightarrow{\text{a.s.}} XY$. Does the corresponding result hold for the other modes of convergence?

3. Let $g : \mathbb{R} \to \mathbb{R}$ be continuous. Show that $g(X_n) \xrightarrow{\mathrm{P}} g(X)$ if $X_n \xrightarrow{\mathrm{P}} X$.

4. Let $Y_1, Y_2, \ldots$ be independent identically distributed variables, each of which can take any value in $\{0, 1, \ldots, 9\}$ with equal probability $\frac{1}{10}$. Let $X_n = \sum_{i=1}^n Y_i\, 10^{-i}$. Show by the use of characteristic functions that X_n converges in distribution to the uniform distribution on $[0, 1]$. Deduce that $X_n \xrightarrow{\text{a.s.}} Y$ for some Y which is uniformly distributed on $[0, 1]$.

5. Let $N(t)$ be a Poisson process with constant intensity on $\mathbb{R}$.

(a) Find the covariance of $N(s)$ and $N(t)$.

(b) Show that N is continuous in mean square, which is to say that $\mathbb{E}\big(\{N(t+h)-N(t)\}^2\big) \to 0$ as $h \to 0$.

(c) Prove that N is continuous in probability, which is to say that $\mathbb{P}\big(|N(t+h)-N(t)| > \epsilon\big) \to 0$ as $h \to 0$, for all $\epsilon > 0$.

(d) Show that N is differentiable in probability but not in mean square.

6. Prove that $n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} 0$ whenever the X_i are independent identically distributed variables with zero means and such that $\mathbb{E}(X_1^4) < \infty$.

7. Show that $X_n \xrightarrow{\text{a.s.}} X$ whenever $\sum_n \mathbb{E}(|X_n - X|^r) < \infty$ for some $r > 0$.

8. Show that if $X_n \xrightarrow{\text{D}} X$ then $aX_n + b \xrightarrow{\text{D}} aX + b$ for any real a and b.

9. If X has zero mean and variance σ^2, show that

$$\mathbb{P}(X \geq t) \leq \frac{\sigma^2}{\sigma^2 + t^2} \quad \text{for } t > 0.$$

10. Show that $X_n \xrightarrow{\text{P}} 0$ if and only if

$$\mathbb{E}\left(\frac{|X_n|}{1+|X_n|}\right) \to 0 \quad \text{as } n \to \infty.$$

11. The sequence $\{X_n\}$ is said to be *mean-square Cauchy convergent* if $\mathbb{E}\{(X_n - X_m)^2\} \to 0$ as $m, n \to \infty$. Show that $\{X_n\}$ converges in mean square to some limit X if and only if it is mean-square Cauchy convergent. Does the corresponding result hold for the other modes of convergence?

12. Suppose that $\{X_n\}$ is a sequence of uncorrelated variables with zero means and uniformly bounded variances. Show that $n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{\text{m.s.}} 0$.

13. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with the common distribution function F, and suppose that $F(x) < 1$ for all x. Let $M_n = \max\{X_1, X_2, \ldots, X_n\}$ and suppose that there exists a strictly increasing unbounded positive sequence $a_1, a_2, \ldots$ such that $\mathbb{P}(M_n/a_n \leq x) \to H(x)$ for some distribution function H. Let us assume that H is continuous with $0 < H(1) < 1$; substantially weaker conditions suffice but introduce extra difficulties.

(a) Show that $n[1 - F(a_n x)] \to -\log H(x)$ as $n \to \infty$ and deduce that

$$\frac{1-F(a_n x)}{1-F(a_n)} \to \frac{\log H(x)}{\log H(1)} \quad \text{if } x > 0.$$

(b) Deduce that if $x > 0$

$$\frac{1-F(tx)}{1-F(t)} \to \frac{\log H(x)}{\log H(1)} \quad \text{as } t \to \infty.$$

(c) Set $x = x_1 x_2$ and make the substitution

$$g(x) = \frac{\log H(e^x)}{\log H(1)}$$

to find that $g(x+y) = g(x)g(y)$, and deduce that

$$H(x) = \begin{cases} \exp(-\alpha x^{-\beta}) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

for some non-negative constants α and β.

You have shown that H is the distribution function of Y^{-1}, where Y has a Weibull distribution.

14. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with the Cauchy distribution. Show that $M_n = \max\{X_1, X_2, \ldots, X_n\}$ is such that $\pi M_n/n$ converges in distribution, the limiting distribution function being given by $H(x) = e^{-1/x}$ if $x \geq 0$.

15. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables whose common characteristic function ϕ satisfies $\phi'(0) = i\mu$. Show that $n^{-1}\sum_{j=1}^{n} X_j \xrightarrow{\mathrm{P}} \mu$.

16. Total variation distance. The *total variation distance* $d_{\mathrm{TV}}(X, Y)$ between two random variables X and Y is defined by

$$d_{\mathrm{TV}}(X, Y) = \sup_{u:\|u\|_\infty=1} \big|\mathbb{E}(u(X)) - \mathbb{E}(u(Y))\big|$$

where the supremum is over all (measurable) functions $u : \mathbb{R} \to \mathbb{R}$ such that $\|u\|_\infty = \sup_x |u(x)|$ satisfies $\|u\|_\infty = 1$.

(a) If X and Y are discrete with respective masses f_n and g_n at the points x_n, show that

$$d_{\mathrm{TV}}(X, Y) = \sum_n |f_n - g_n| = 2 \sup_{A \subseteq \mathbb{R}} \big|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\big|.$$

(b) If X and Y are continuous with respective density functions f and g, show that

$$d_{\mathrm{TV}}(X, Y) = \int_{-\infty}^{\infty} |f(x) - g(x)|\, dx = 2 \sup_{A \subseteq \mathbb{R}} \big|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\big|.$$

(c) Show that $d_{\mathrm{TV}}(X_n, X) \to 0$ implies that $X_n \to X$ in distribution, but that the converse is false.

(d) **Maximal coupling.** Show that $\mathbb{P}(X \neq Y) \geq \frac{1}{2} d_{\mathrm{TV}}(X, Y)$, and that there exists a pair X', Y' having the same marginals for which equality holds.

(e) If X_i, Y_j are independent random variables, show that

$$d_{\mathrm{TV}}\Big(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\Big) \leq \sum_{i=1}^{n} d_{\mathrm{TV}}(X_i, Y_i).$$

17. Let $g : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Show that

$$\sum_{k=0}^{\infty} g(k/n) \frac{(n\lambda)^k}{k!} e^{-n\lambda} \to g(\lambda) \quad \text{as } n \to \infty.$$

18. Let X_n and Y_m be independent random variables having the Poisson distribution with parameters n and m, respectively. Show that

$$\frac{(X_n - n) - (Y_m - m)}{\sqrt{X_n + Y_m}} \xrightarrow{\mathrm{D}} N(0, 1) \quad \text{as } m, n \to \infty.$$

19. (a) Suppose that $X_1, X_2, \ldots$ is a sequence of random variables, each having a normal distribution, and such that $X_n \xrightarrow{\mathrm{D}} X$. Show that X has a normal distribution, possibly degenerate.

(b) For each $n \geq 1$, let (X_n, Y_n) be a pair of random variables having a bivariate normal distribution. Suppose that $X_n \xrightarrow{\mathrm{P}} X$ and $Y_n \xrightarrow{\mathrm{P}} Y$, and show that the pair (X, Y) has a bivariate normal distribution.

20. Let $X_1, X_2, \ldots$ be random variables satisfying $\text{var}(X_n) < c$ for all n and some constant c. Show that the sequence obeys the weak law, in the sense that $n^{-1}\sum_1^n (X_i - \mathbb{E}X_i)$ converges in probability to 0, if the correlation coefficients satisfy either of the following:

(i) $\rho(X_i, X_j) \le 0$ for all $i \ne j$,

(ii) $\rho(X_i, X_j) \to 0$ as $|i - j| \to \infty$.

21. Let $X_1, X_2, \ldots$ be independent random variables with common density function

$$f(x) = \begin{cases} 0 & \text{if } |x| \le 2, \\ \dfrac{c}{x^2 \log |x|} & \text{if } |x| > 2, \end{cases}$$

where c is a constant. Show that the X_i have no mean, but $n^{-1}\sum_{i=1}^n X_i \xrightarrow{\text{P}} 0$ as $n \to \infty$. Show that convergence does not take place almost surely.

22. Let X_n be the Euclidean distance between two points chosen independently and uniformly from the n-dimensional unit cube. Show that $\mathbb{E}(X_n)/\sqrt{n} \to 1/\sqrt{6}$ as $n \to \infty$.

23. Let $X_1, X_2, \ldots$ be independent random variables having the uniform distribution on $[-1, 1]$. Show that

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i^{-1}\right| > \tfrac{1}{2}n\pi\right) \to \tfrac{1}{2} \quad \text{as } n \to \infty.$$

24. Let $X_1, X_2, \ldots$ be independent random variables, each X_k having mass function given by

$$\mathbb{P}(X_k = k) = \mathbb{P}(X_k = -k) = \frac{1}{2k^2},$$
$$\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = \frac{1}{2}\left(1 - \frac{1}{k^2}\right) \qquad \text{if } k > 1.$$

Show that $U_n = \sum_1^n X_i$ satisfies $U_n/\sqrt{n} \xrightarrow{\text{D}} N(0, 1)$ but $\text{var}(U_n/\sqrt{n}) \to 2$ as $n \to \infty$.

25. Let $X_1, X_2, \ldots$ be random variables, and let $N_1, N_2, \ldots$ be random variables taking values in the positive integers such that $N_k \xrightarrow{\text{P}} \infty$ as $k \to \infty$. Show that:

(i) if $X_n \xrightarrow{\text{D}} X$ and the X_n are independent of the N_k, then $X_{N_k} \xrightarrow{\text{D}} X$ as $k \to \infty$,

(ii) if $X_n \xrightarrow{\text{a.s.}} X$ then $X_{N_k} \xrightarrow{\text{P}} X$ as $k \to \infty$.

26. Stirling's formula.

(a) Let $a(k, n) = n^k/(k-1)!$ for $1 \le k \le n+1$. Use the fact that $1 - x \le e^{-x}$ if $x \ge 0$ to show that

$$\frac{a(n-k, n)}{a(n+1, n)} \le e^{-k^2/(2n)} \qquad \text{if } k \ge 0.$$

(b) Let $X_1, X_2, \ldots$ be independent Poisson variables with parameter 1, and let $S_n = X_1 + \cdots + X_n$. Define the function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(x) = \begin{cases} -x & \text{if } 0 \ge x \ge -M, \\ 0 & \text{otherwise,} \end{cases}$$

where M is large and positive. Show that, for large n,

$$\mathbb{E}\left(g\left\{\frac{S_n - n}{\sqrt{n}}\right\}\right) = \frac{e^{-n}}{\sqrt{n}}\{a(n+1, n) - a(n-k, n)\}$$

where $k = \lfloor Mn^{1/2} \rfloor$. Now use the central limit theorem and (a) above, to deduce Stirling's formula:

$$\frac{n!\, e^n}{n^{n+\frac{1}{2}}\sqrt{2\pi}} \to 1 \qquad \text{as } n \to \infty.$$

27. A bag contains red and green balls. A ball is drawn from the bag, its colour noted, and then it is returned to the bag together with a new ball of the same colour. Initially the bag contained one ball of each colour. If R_n denotes the number of red balls in the bag after n additions, show that $S_n = R_n/(n+2)$ is a martingale. Deduce that the ratio of red to green balls converges almost surely to some limit as $n \to \infty$.

28. Anscombe's theorem. Let $\{X_i : i \geq 1\}$ be independent identically distributed random variables with zero mean and finite positive variance σ^2, and let $S_n = \sum_1^n X_i$. Suppose that the integer-valued random process $M(t)$ satisfies $t^{-1}M(t) \xrightarrow{\text{P}} \theta$ as $t \to \infty$, where θ is a positive constant. Show that

$$\frac{S_{M(t)}}{\sigma\sqrt{\theta t}} \xrightarrow{\text{D}} N(0,1) \quad \text{and} \quad \frac{S_{M(t)}}{\sigma\sqrt{M(t)}} \xrightarrow{\text{D}} N(0,1) \quad \text{as } t \to \infty.$$

You should not assume that the process M is independent of the X_i.

29. Kolmogorov's inequality. Let $X_1, X_2, \ldots$ be independent random variables with zero means, and $S_n = X_1 + X_2 + \cdots + X_n$. Let $M_n = \max_{1 \leq k \leq n} |S_k|$ and show that $\mathbb{E}(S_n^2 I_{A_k}) > c^2 \mathbb{P}(A_k)$ where $A_k = \{M_{k-1} \leq c < M_k\}$ and $c > 0$. Deduce Kolmogorov's inequality:

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| > c\right) \leq \frac{\mathbb{E}(S_n^2)}{c^2}, \qquad c > 0.$$

30. Let $X_1, X_2, \ldots$ be independent random variables with zero means, and let $S_n = X_1 + X_2 + \cdots + X_n$. Using Kolmogorov's inequality or the martingale convergence theorem, show that:

(i) $\sum_{i=1}^{\infty} X_i$ converges almost surely if $\sum_{k=1}^{\infty} \mathbb{E}(X_k^2) < \infty$,

(ii) if there exists an increasing real sequence (b_n) such that $b_n \to \infty$, and satisfying the inequality $\sum_{k=1}^{\infty} \mathbb{E}(X_k^2)/b_k^2 < \infty$, then $b_n^{-1} \sum_{k=1}^{\infty} X_k \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

31. Estimating the transition matrix. The Markov chain $X_0, X_1, \ldots, X_n$ has initial distribution $f_i = \mathbb{P}(X_0 = i)$ and transition matrix $\mathbf{P}$. The *log-likelihood* function $\lambda(\mathbf{P})$ is defined as $\lambda(\mathbf{P}) = \log(f_{X_0} p_{X_0,X_1} p_{X_1,X_2} \cdots p_{X_{n-1},X_n})$. Show that:

(a) $\lambda(\mathbf{P}) = \log f_{X_0} + \sum_{i,j} N_{ij} \log p_{ij}$ where N_{ij} is the number of transitions from i to j,

(b) viewed as a function of the p_{ij}, $\lambda(\mathbf{P})$ is maximal when $p_{ij} = \widehat{p}_{ij}$ where $\widehat{p}_{ij} = N_{ij}/\sum_k N_{ik}$,

(c) if X is irreducible and ergodic then $\widehat{p}_{ij} \xrightarrow{\text{a.s.}} p_{ij}$ as $n \to \infty$.

32. Ergodic theorem in discrete time. Let X be an irreducible discrete-time Markov chain, and let μ_i be the mean recurrence time of state i. Let $V_i(n) = \sum_{r=0}^{n-1} I_{\{X_r = i\}}$ be the number of visits to i up to $n-1$, and let f be any bounded function on S. Show that:

(a) $n^{-1}V_i(n) \xrightarrow{\text{a.s.}} \mu_i^{-1}$ as $n \to \infty$,

(b) if $\mu_i < \infty$ for all i, then

$$\frac{1}{n} \sum_{r=0}^{n-1} f(X_r) \to \sum_{i \in S} f(i)/\mu_i \quad \text{as } n \to \infty.$$

33. Ergodic theorem in continuous time. Let X be an irreducible persistent continuous-time Markov chain with generator $\mathbf{G}$ and finite mean recurrence times μ_j.

(a) Show that $\frac{1}{t} \int_0^t I_{\{X(s)=j\}}\, ds \xrightarrow{\text{a.s.}} \frac{1}{\mu_j g_j}$ as $t \to \infty$;

(b) deduce that the stationary distribution $\boldsymbol{\pi}$ satisfies $\pi_j = 1/(\mu_j g_j)$;
(c) show that, if f is a bounded function on S,

$$\frac{1}{t}\int_0^t f(X(s))\,ds \xrightarrow{\text{a.s.}} \sum_i \pi_i f(i) \quad \text{as } t \to \infty.$$

34. Tail equivalence. Suppose that the sequences $\{X_n : n \geq 1\}$ and $\{Y_n : n \geq 1\}$ are *tail equivalent*, which is to say that $\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) < \infty$. Show that:
(a) $\sum_{n=1}^{\infty} X_n$ and $\sum_{n=1}^{\infty} Y_n$ converge or diverge together,
(b) $\sum_{n=1}^{\infty}(X_n - Y_n)$ converges almost surely,
(c) if there exist a random variable X and a sequence a_n such that $a_n \uparrow \infty$ and $a_n^{-1}\sum_{r=1}^{n} X_r \xrightarrow{\text{a.s.}} X$, then

$$\frac{1}{a_n}\sum_{r=1}^{n} Y_r \xrightarrow{\text{a.s.}} X.$$

35. Three series theorem. Let $\{X_n : n \geq 1\}$ be independent random variables. Show that $\sum_{n=1}^{\infty} X_n$ converges a.s. if, for some $a > 0$, the following three series all converge:
(a) $\sum_n \mathbb{P}(|X_n| > a)$,
(b) $\sum_n \operatorname{var}(X_n I_{\{|X_n| \leq a\}})$,
(c) $\sum_n \mathbb{E}(X_n I_{\{|X_n| \leq a\}})$.
[The converse holds also, but is harder to prove.]

36. Let $\{X_n : n \geq 1\}$ be independent random variables with continuous common distribution function F. We call X_k a *record value* for the sequence if $X_k > X_r$ for $1 \leq r < k$, and we write I_k for the indicator function of the event that X_k is a record value.
(a) Show that the random variables I_k are independent.
(b) Show that $R_m = \sum_{k=1}^{m} I_r$ satisfies $R_m/\log m \xrightarrow{\text{a.s.}} 1$ as $m \to \infty$.

37. Random harmonic series. Let $\{X_n : n \geq 1\}$ be a sequence of independent random variables with $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = \frac{1}{2}$. Does the series $\sum_{r=1}^{n} X_r/r$ converge a.s. as $n \to \infty$?