Chapter 10
Probability

This chapter is devoted to an introduction to probability theory. It contains some of the fundamental results of probability theory—the strong law of large numbers, the central limit theorem, the martingale convergence theorem, the construction of Brownian motion processes, and Kolmogorov's consistency theorem.

One purpose of this chapter is to give the reader a chance to work through some applications of measure theory and thereby to get some practice with the techniques presented earlier. Another, perhaps more significant, goal is to give the reader a broader picture of how σ-algebras, measures, measurable functions, and integrals arise.

10.1 Basics

In probability theory one describes and analyzes random situations, often called *experiments*. Let us look at how such situations can be modeled using measure theory. We begin with some terminology.

A *probability space* is a measure space $(\Omega, \mathscr{A}, P)$ such that $P(\Omega) = 1$. The elements of Ω are called the *elementary outcomes* or the *sample points* of our experiment, and the members of $\mathscr{A}$ are called *events*. If $A \in \mathscr{A}$, then $P(A)$ is the *probability* of the event A.

Example 10.1.1. We illustrate these concepts with a very simple example. Suppose we toss a fair coin (one for which a head has probability $1/2$) twice. There are four possible outcomes: we get two heads, we get a head and then a tail, we get a tail and then a head, or we get two tails. So we can let our set Ω of elementary outcomes be $\{HH, HT, TH, TT\}$. It is natural in this case to let $\mathscr{A}$ contain all the subsets of Ω. For example, $\{HT, TH\}$ is one of the subsets of Ω; it corresponds to the real-world event in which we get a head on exactly one of the tosses. Finally, in this situation each elementary outcome has probability $1/4$ of occurring, and so we let the probability of an event A be $1/4$ times the number of elements of A. □

D.L. Cohn, *Measure Theory: Second Edition*, Birkhäuser Advanced Texts Basler Lehrbücher, DOI 10.1007/978-1-4614-6956-8_10,

A *real-valued random variable* on a probability space $(\Omega, \mathscr{A}, P)$ is an $\mathscr{A}$-measurable function from Ω to $\mathbb{R}$. Such a variable represents a numerical observation or measurement whose value depends on the outcome of the random experiment represented by $(\Omega, \mathscr{A}, P)$. More generally, a *random variable* with values in a measurable space $(S, \mathscr{B})$ is a measurable function from $(\Omega, \mathscr{A}, P)$ to $(S, \mathscr{B})$. Let X be a random variable with values in $(S, \mathscr{B})$. The *distribution* of X is the measure PX^{-1} defined on $(S, \mathscr{B})$ by $(PX^{-1})(A) = P(X^{-1}(A))$ (see Sect. 2.6). We will often write P_X for the distribution of a random variable X. If $X_1, \ldots, X_d$ are $(S, \mathscr{B})$-valued random variables on $(\Omega, \mathscr{A}, P)$, then the formula $X(\omega) = (X_1(\omega), \ldots, X_d(\omega))$ defines an S^d-valued random variable X; the distribution of X is called the *joint distribution* of $X_1, \ldots, X_d$.

Example 10.1.2. Let us continue with our coin-tossing example. The number of heads that appear when our two coins are tossed can be represented with the random variable X defined by

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = TT, \\ 1 & \text{if } \omega = HT \text{ or } \omega = TH, \text{ and} \\ 2 & \text{if } \omega = HH. \end{cases}$$

The distribution P_X of X is given by $P_X = \frac{1}{4}\delta_0 + \frac{1}{2}\delta_1 + \frac{1}{4}\delta_2$. □

An abbreviated notation for events is common in probability. We introduce it with a couple of examples. Suppose that $(\Omega, \mathscr{A}, P)$ is a probability space and that X and X_n, $n = 1, 2, \ldots$, are real-valued random variables on Ω. Then the events

$$\{\omega \in \Omega : X(\omega) \geq 0\},$$

$$\{\omega \in \Omega : X(\omega) = \lim_n X_n(\omega)\},$$

and

$$\{\omega \in \Omega : \lim_n X_n(\omega) \text{ exists}\}$$

are often abbreviated as $\{X \geq 0\}$, $\{X = \lim_n X_n\}$, and $\{\lim_n X_n \text{ exists}\}$. Sometimes one goes a bit further and simply writes $P(X \geq 0)$ instead of $P(\{X \geq 0\})$ or $P(\{\omega \in \Omega : X(\omega) \geq 0\})$.

If a real-valued random variable X on a probability space $(\Omega, \mathscr{A}, P)$ is integrable with respect to P, then its *expected value*, or *expectation*, written $E(X)$, is the integral of X with respect to P. That is, $E(X) = \int X\, dP$. If X is integrable, one also says that X *has a finite expected value* or that X *has an expected value*. Note that Proposition 2.6.8 gives a way to compute the expected value of a real-valued random variable in terms of its distribution, namely $E(X) = \int_{\mathbb{R}} x P_X(dx)$. That proposition in fact gives the more general formula $E(f \circ X) = \int_{\mathbb{R}} f\, dP_X$, by which we can compute the expected value of a Borel function f of a random variable X in terms of the distribution of X.

We often have use for the expected value of the square of a real-valued random variable X, or the *second moment* of X. If X has a finite second moment, then it follows from the inequality $|X| \le X^2 + 1$ that X has a finite expectation. In this case, one calls the expected value of $(X - E(X))^2$ the *variance* of X; it gives a measure of the amount by which the values of X differ from the expected value of X. The nonnegative square root of the variance of X is called the *standard deviation* of X. One often denotes the expected value of a random variable X with μ_X or simply μ, the variance with $\operatorname{var}(X)$ or σ_X^2, and the standard deviation with σ_X.

Lemma 10.1.3. *Let X be a random variable with a finite second moment, and let a and b be real numbers. Then*

(a) $\operatorname{var}(X) = E(X^2) - (E(X))^2$, *and*
(b) $\operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$.

Proof. The lemma follows from basic algebra and the linearity of the integral. □

Suppose that X is a real-valued random variable with a discrete distribution—that is, suppose that there is a countable subset C of $\mathbb{R}$ such that $P(X \in C) = 1$. Then X has a finite expected value if and only if $\sum_{x \in C} |x| P(X = x) < +\infty$, and in that case $E(X) = \sum_{x \in C} x P(X = x)$. Likewise, if the distribution P_X of X is absolutely continuous with respect to Lebesgue measure and if f_X is the Radon–Nikodym derivative of P_X with respect to Lebesgue measure, then X has a finite expected value if and only if $\int_{\mathbb{R}} |x| f_X(x)\, dx < +\infty$, and in that case $E(X) = \int_{\mathbb{R}} x f_X(x)\, dx$. As these remarks may suggest, it turns out that discrete and continuous random variables,[1] which receive separate treatments in elementary discussions of probability theory, can be given a fairly uniform treatment in terms of measure theory.

We have seen (in Propositions 1.3.9 and 1.3.10) that there is a correspondence between finite Borel measures on $\mathbb{R}$ and bounded nondecreasing right-continuous functions $F : \mathbb{R} \to \mathbb{R}$ for which $\lim_{x \to -\infty} F(x) = 0$. In the present context, this means that the distribution P_X of a real-valued random variable X is determined by the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by

$$F_X(x) = P_X((-\infty, x]) = P(X \le x).$$

The function F_X is called the *cumulative distribution function*, or just the *distribution function*, of X.

Let $\{X_i\}_{i \in I}$ be an indexed family of random variables on a probability space $(\Omega, \mathscr{A}, P)$. Then $\sigma(X_i, i \in I)$ is the smallest σ-algebra on Ω that makes all these variables measurable. Likewise, if $\{X_n\}$ is a sequence of random variables on $(\Omega, \mathscr{A}, P)$, then one often writes $\sigma(X_1, X_2, \dots)$ for the smallest σ-algebra on Ω that makes each X_n measurable.

[1] A real-valued random variable is *discrete* if its distribution is discrete and is *continuous* if its distribution is absolutely continuous with respect to Lebesgue measure.

Examples 10.1.4.

(a) We begin by returning to coin tossing. Suppose that now our experiment is to toss a fair coin repeatedly, until we first get a head, and then to stop. It seems reasonable to define Ω by

$$\Omega = \{H, TH, TTH, \ldots, TTTTTTTH, \ldots\}$$

and to let $\mathscr{A}$ consist of all subsets of Ω. We will (by countable additivity) determine the probability of all the events in $\mathscr{A}$ if we specify the probabilities of the one-point subsets of Ω. It seems reasonable to let $P(\{H\}) = 1/2$, $P(\{TH\}) = 1/4$, $P(\{TTH\}) = 1/8$, ... (the reader should think through this assignment of probabilities again, after reading the discussion of independence that occurs later in this section). Note that the sum of the geometric series $\sum_{n=1}^{\infty}(1/2)^n$ is 1, and so this assignment of probabilities does give a probability measure.

(b) Now suppose that we choose a real number from the interval $[a,b]$ in such a way that the probability that the number chosen lies in a subinterval I of $[a,b]$ is proportional to the length of I. We can describe this situation with the probability space $([a,b],\mathscr{B}([a,b]),P)$, where the measure P is given by $P(A) = \lambda(A)/(b-a)$. In this case one has a *uniform distribution* on $[a,b]$. Of course, if the interval $[a,b]$ is the unit interval $[0,1]$, then the measure P is just the restriction of Lebesgue measure to the Borel subsets of $[0,1]$.

(c) Now suppose that f is a nonnegative Borel measurable function on $\mathbb{R}$ such that $\int f\,d\lambda = 1$. Then the formula $P(A) = \int_A f\,d\lambda$ defines a probability measure on the measurable space $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. The function f is called the *density* of P (or of a random variable having distribution P). Note that the measures in part (b) above can be viewed as special cases of the situation here, with the uniform distribution on $[a,b]$ given by the density function that has value $1/(b-a)$ on $[a,b]$ and 0 elsewhere.

(d) In a similar way, a nonnegative Borel measurable function on $\mathbb{R}^2$ such that $\int\int f(x,y)\,\lambda(dx)\lambda(dy) = 1$ defines a probability measure on the measurable space $(\mathbb{R}^2,\mathscr{B}(\mathbb{R}^2))$.

(e) Let us now look at *normal*, or *Gaussian*, distributions, which are given by the familiar bell-shaped curves. We begin by evaluating the integral $\int_{\mathbb{R}} e^{-x^2/2}\,dx$. Let us denote the value of this integral by A for a moment. If we interpret A^2 as an integral over $\mathbb{R}^2$ and evaluate the integral using polar coordinates, we find

$$A^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy = \int_0^{2\pi}\int_0^{\infty} re^{-r^2/2}\,dr\,d\theta = 2\pi.$$

Thus $A = \sqrt{2\pi}$, and so the function $x \mapsto \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ is a probability density function on $\mathbb{R}$ (that is, it is nonnegative and its integral over $\mathbb{R}$ is 1).

Now suppose that X is a random variable whose distribution has density $x \mapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. It is easy to check that

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x e^{-x^2/2}\, dx = 0$$

and hence that $E(X) = 0$. If in the following calculation we use integration by parts to convert the first integral into the second, whose value we know, we find that

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2}\, dx = 1$$

and hence that $E(X^2) = 1$. Thus X has expected value 0 and variance 1.

It is easy to check that if X is as above and if μ and σ are constants, with $\sigma > 0$, then the random variable $\sigma X + \mu$ has mean μ and variance σ^2 (see Lemma 10.1.3). Furthermore, according to Lemma 10.1.5, $\sigma X + \mu$ has density g_{μ,σ^2} given by

$$g_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

With this we have the densities of the *normal* or *Gaussian* random variables with mean μ and variance σ^2.

One often writes $N(0,1)$ for the distribution of a normal random variable with mean 0 and variance 1 and $N(\mu,\sigma^2)$ for the distribution of a normal random variable with mean μ and variance σ^2. Thus $N(0,1)$ is the measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with density $x \mapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and $N(\mu,\sigma^2)$ is the measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with density g_{μ,σ^2}. □

Lemma 10.1.5. *Let X be a real-valued random variable with density f_X, let a and b be real constants with $a > 0$, and let $Y = aX + b$. Then Y is a continuous random variable whose density f_Y is given by*

$$f_Y(t) = \frac{1}{a} f_X\left(\frac{t-b}{a}\right).$$

Proof. Define a function $T\colon \mathbb{R} \to \mathbb{R}$ by $T(t) = at + b$. Then $\lambda(T(A)) = a\lambda(A)$ holds for each subinterval A of $\mathbb{R}$ and consequently for each Borel subset A of $\mathbb{R}$. Thus

$$a \int h\, d\lambda = \int h \circ T^{-1}\, d\lambda$$

holds for each nonnegative measurable h (check this first in the case where h is the indicator function of a Borel set), and so we have

$$P_Y(A) = P_X(T^{-1}(A)) = \int_{T^{-1}(A)} f_X \, d\lambda$$
$$= (1/a) \int (\chi_{T^{-1}(A)} \circ T^{-1})(f_X \circ T^{-1}) \, d\lambda$$
$$= (1/a) \int_A f_X \left(\frac{t-b}{a} \right) \lambda(dt).$$

Thus P_Y can be calculated by integrating the function $t \mapsto \frac{1}{a} f_X(\frac{t-b}{a})$, and the proof is complete. □

We will need the following fact about normal distributions.

Lemma 10.1.6. *Let Z be a normal random variable with mean 0 and variance 1. Then*

$$P(Z \geq A) \leq \frac{1}{\sqrt{2\pi}A} e^{-A^2/2}$$

holds for each positive real number A.

Proof. We have

$$P(Z \geq A) = \frac{1}{\sqrt{2\pi}} \int_A^\infty e^{-x^2/2} \, dx \leq \frac{1}{\sqrt{2\pi}} \int_A^\infty \frac{x}{A} e^{-x^2/2} \, dx = \frac{1}{\sqrt{2\pi}A} e^{-A^2/2}.$$

□

Let us turn to a few definitions and results involving independence.

Let $(\Omega, \mathscr{A}, P)$ be a probability space, and let $\{A_i\}_{i \in I}$ be an indexed family of events. The events[2] A_i, $i \in I$, are called *independent* if for each finite subset I_0 of I we have

$$P(\cap_{i \in I_0} A_i) = \prod_{i \in I_0} P(A_i).$$

Let $\{X_i\}_{i \in I}$ be an indexed family of random variables, defined on $(\Omega, \mathscr{A}, P)$ and with values in the measurable space $(S, \mathscr{B})$. The random variables X_i, $i \in I$, are called *independent* if for each choice of sets A_i in $\mathscr{B}$, $i \in I$, the events $X_i^{-1}(A_i)$ are independent.

Finally, let $(\Omega, \mathscr{A}, P)$ be a probability space and let $\{\mathscr{A}_i\}_{i \in I}$ be an indexed family of sub-σ-algebras of $\mathscr{A}$. The σ-algebras $\mathscr{A}_i$, $i \in I$, are *independent* if for each choice of sets A_i in $\mathscr{A}_i$, $i \in I$, the events A_i are independent.

Note that if $\{X_i\}_{i \in I}$ is an indexed family of random variables on a probability space $(\Omega, \mathscr{A}, P)$, then the random variables X_i, $i \in I$, are independent if and only if the σ-algebras $\sigma(X_i)$, $i \in I$, are independent.

[2] Although the independence of A_i, $i \in I$, depends on the relationship between the events A_i, rather than on the events individually, it is standard to call the events, rather than the indexed family, independent.

Proposition 10.1.7. *Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\{\mathscr{A}_i\}_{i\in I}$ be an indexed family of independent sub-σ-algebras of $\mathscr{A}$, let $\{S_j\}_{j\in J}$ be a partition of I, and for each j in J let $\mathscr{B}_j = \sigma(\cup_{i\in S_j}\mathscr{A}_i)$. Then the σ-algebras $\mathscr{B}_j$ are independent.*

Proof. For each j in J let $\mathscr{P}_j$ consist of all finite intersections of sets in $\cup_{i\in S_j}\mathscr{A}_i$. Note that each $\mathscr{P}_j$ is a π-system such that $\mathscr{B}_j = \sigma(\mathscr{P}_j)$. Let J_0 be a nonempty finite subset of J, and for each j in J_0 let A_j be a member of $\mathscr{P}_j$. The relation

$$P(\cap_{j\in J_0} A_j) = \prod_{j\in J_0} P(A_j) \tag{1}$$

follows from the independence of the $\mathscr{A}_i$'s. Now suppose that the elements of J_0 are $j_1, j_2, \ldots, j_n$, and let $\mathscr{D}$ be the class of all A in $\mathscr{B}_{j_n}$ such that

$$P(A_{j_1} \cap \cdots \cap A_{j_{n-1}} \cap A) = P(A_{j_1}) \cdots P(A_{j_{n-1}}) P(A)$$

holds for all A_{j_i} in $\mathscr{P}_{j_i}$, $i = 1, \ldots, n-1$. Then $\mathscr{D}$ is a Dynkin class (i.e., a d-system) that includes $\mathscr{P}_{j_n}$, and so Theorem 1.6.2 implies that $\mathscr{D} = \mathscr{B}_{j_n}$. Similar arguments, $n-1$ of them, show that (1) holds for all A_j in $\mathscr{B}_j$, $j \in J_0$. Since the independence of the $\mathscr{B}_j$, $j \in J$ depends only on the independence of finite subfamilies, the proof is complete. □

Example 10.1.8. Proposition 10.1.7 may look overly abstract, but it allows simple proofs of some results for which a rigorous proof might otherwise be awkward. For example, suppose that $\{X_n\}_{n=1}^{\infty}$ is a sequence of independent random variables on a probability space $(\Omega, \mathscr{A}, P)$. Then it is an immediate consequence of Proposition 10.1.7 that the random variables $X_{2i-1} + X_{2i}$, $i = 1, 2, \ldots$ are independent. Proving this independence in other ways would probably take more work. □

Proposition 10.1.9. *Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $(S, \mathscr{B})$ be a measurable space, let $X_1, X_2, \ldots, X_d$ be S-valued random variables on Ω, and let X be the S^d-valued random variable with components $X_1, X_2, \ldots, X_d$. Let $P_{X_1}, P_{X_2}, \ldots, P_{X_d}$, and P_X be the distributions of $X_1, X_2, \ldots, X_d$, and X, respectively. Then $X_1, X_2, \ldots, X_d$ are independent if and only if the joint distribution P_X is equal to the product measure $P_{X_1} \times P_{X_2} \times \cdots \times P_{X_d}$.*

Proof. If we rewrite the definition of independence, we find that $X_1, \ldots, X_d$ are independent if and only if

$$P_X(A_1 \times \cdots \times A_d) = \prod_i P_{X_i}(A_i)$$

holds for each choice of sets A_i in $\mathscr{B}$, $i = 1, \ldots, d$. Thus if P_X is equal to the product of the measures P_{X_i}, then $X_1, X_2, \ldots, X_d$ are independent. The converse follows from the uniqueness of product measures (see Theorem 5.1.4 and the discussion at the end of Sect. 5.2). □

Proposition 10.1.10. *Let $(\Omega, \mathscr{A}, P)$ be a probability space and let $X_1, X_2, \ldots, X_n$ be independent real-valued random variables on $(\Omega, \mathscr{A}, P)$, each of which has a finite expectation. Then the product $\prod_i X_i$ has a finite expectation, and $E(\prod_i X_i) = \prod_i E(X_i)$.*

Proof. Let X be the $\mathbb{R}^n$-valued random variable with components $X_1, \ldots, X_n$, and let P_X and $P_{X_1}, \ldots, P_{X_n}$ be the distributions of X and $X_1, \ldots, X_n$. We will use these distributions for the calculation of $E(\prod_i X_i)$ and $\prod_i E(X_i)$. Since the random variables $X_i, \ldots, X_n$ are independent, P_X is the product of the measures $P_{X_1}, \ldots, P_{X_n}$ (Proposition 10.1.9). Thus we can use Proposition 5.2.1 and Theorem 5.2.2, together with the finiteness of the expectations $E(X_i)$ and the remarks at the end of Sect. 5.2, to conclude that $\prod_i X_i$ has a finite expectation and that $E(\prod_i X_i) = \prod_i E(X_i)$. □

Corollary 10.1.11. *Let $X_1, X_2, \ldots, X_n$ be independent real-valued random variables with finite second moments, and let $S = X_1 + \cdots + X_n$. Then $\operatorname{var}(S) = \sum_i \operatorname{var}(X_i)$.*

Proof. By the independence of X_i and X_j (where $i \neq j$), the expectation of the product $(X_i - E(X_i))(X_j - E(X_j))$ is the product of the expectations of $X_i - E(X_i)$ and $X_j - E(X_j)$, namely 0. Thus

$$\begin{aligned}\operatorname{var}(S) &= E\big(\big(\sum_i (X_i - E(X_i))\big)^2\big) = \sum_i \sum_j E\big((X_i - E(X_i))(X_j - E(X_j))\big) \\ &= \sum_i E\big((X_i - E(X_i))^2\big) = \sum_i \operatorname{var}(X_i).\end{aligned}$$

□

Now suppose that X_1 and X_2 are independent real-valued (or $\mathbb{R}^d$-valued) random variables with distributions P_{X_1} and P_{X_2}. In view of Proposition 10.1.9, we can use the product measure $P_{X_1} \times P_{X_2}$ to compute the distribution $P_{X_1+X_2}$ of $X_1 + X_2$:

$$P_{X_1+X_2}(A) = (P_{X_1} \times P_{X_2})(\{(x_1, x_2) : x_1 + x_2 \in A\}). \tag{2}$$

One defines the *convolution* $\nu_1 * \nu_2$ of finite measures ν_1 and ν_2 on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ by

$$(\nu_1 * \nu_2)(A) = (\nu_1 \times \nu_2)(\{(x_1, x_2) : x_1 + x_2 \in A\});$$

thus (2) says that the distribution of the sum of two independent random variables is the convolution of their distributions: $P_{X_1+X_2} = P_{X_1} * P_{X_2}$.

Note that convolution satisfies the associative law $\nu_1 * (\nu_2 * \nu_3) = (\nu_1 * \nu_2) * \nu_3$, since if X_1, X_2, and X_3 are independent random variables with distributions ν_1, ν_2, and ν_3, then both $\nu_1 * (\nu_2 * \nu_3)$ and $(\nu_1 * \nu_2) * \nu_3$ give the distribution of $X_1 + X_2 + X_3$. More generally, the convolution of the distributions of n independent random variables gives the distribution of their sum.

We can compute convolutions as follows.

Proposition 10.1.12. *Let ν_1 and ν_2 be probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.*

(a) *The convolution $\nu_1 * \nu_2$ satisfies*

$$(\nu_1 * \nu_2)(A) = \int \nu_1(A-y)\,d\nu_2(y) = \int \nu_2(A-x)\,d\nu_1(x)$$

for each A in $\mathscr{B}(\mathbb{R}^d)$.

(b) *If ν_1 is absolutely continuous (with respect to Lebesgue measure), with density f, then $\nu_1 * \nu_2$ is absolutely continuous, with density $x \mapsto \int f(x-y)\,\nu_2(dy)$.*

(c) *If ν_1 and ν_2 are absolutely continuous, with densities f and g, then $\nu_1 * \nu_2$ is absolutely continuous, with density $x \mapsto \int f(x-y)g(y)\,\lambda(dy)$.*

Proof. Since the sections of the set $\{(x,y) : x+y \in A\}$ are equal to $A-x$ and $A-y$, part (a) is an immediate consequence of Theorem 5.1.4. Part (b) follows from the calculation

$$\begin{aligned}\nu_1 * \nu_2(A) &= \int\int \chi_A(x+y) f(x)\,\lambda(dx)\,\nu_2(dy)\\ &= \int\int \chi_A(x) f(x-y)\,\lambda(dx)\,\nu_2(dy)\\ &= \int \chi_A(x) \int f(x-y)\,\nu_2(dy)\,\lambda(dx).\end{aligned}$$

(The finiteness of $\int f(x-y)\,\nu_2(dy)$ for almost every x follows from this calculation, applied in the case where $A = \mathbb{R}$.) Part (c) follows from part (b), since in this case we have $\int f(x-y)\,\nu_2(dy) = \int f(x-y)g(y)\,\lambda(dy)$ (recall Exercise 4.2.3). □

In the remainder of this section we look at some random variables that arise when we consider the binary expansions of the values of certain uniformly distributed random variables. The techniques discussed here will give us a way to construct arbitrary sequences of independent (real-valued) random variables.

It will be convenient to have a bit of standard terminology. A random variable X is said to have a *Bernoulli distribution with parameter p* if the possible values[3] of X are 0 and 1, with 1 having probability p and 0 having probability $1-p$.

So let us suppose that $(\Omega, \mathscr{A}, P)$ is a probability space and that X is a random variable on $(\Omega, \mathscr{A}, P)$ that is uniformly distributed on $[0,1]$. By redefining X on a null set, if necessary, we can assume that *every* value of X belongs to $[0,1)$. Define a sequence $\{Y_n\}$ on $(\Omega, \mathscr{A}, P)$ by letting $Y_n(\omega)$ be the nth bit in the binary expansion[4]

[3] Actually, we are only assuming that $P(X \in \{0,1\}) = 1$ and not that $X(\omega) \in \{0,1\}$ for *every* ω in Ω.

[4] In case the value $X(\omega)$ has two binary expansions, take the one that ends in an infinite sequence of 0's. See B.9 in Appendix B.

of $X(\omega)$. Then $Y_1(\omega)$ is 0 if $X(\omega)$ belongs to the interval $[0,1/2)$ and is 1 if $X(\omega)$ belongs to $[1/2,1)$. Likewise $Y_2(\omega)$ is 0 if $X(\omega)$ belongs to $[0,1/4)\cup[1/2,3/4)$ and is 1 if $X(\omega)$ belongs to $[1/4,1/2)\cup[3/4,1)$. In general, $Y_n(\omega)$ is 0 if $X(\omega)$ satisfies $2i/2^n \le X(\omega) < (2i+1)/2^n$ for some i and is 1 otherwise; from that it is not difficult to check that the variables $\{Y_n\}$ are measurable and independent, with each having a Bernoulli distribution with parameter $1/2$.

Proposition 10.1.13. *Let $(\Omega,\mathscr{A},P)$ be a probability space.*

(a) *Suppose that X is a random variable on $(\Omega,\mathscr{A},P)$ that is uniformly distributed on $[0,1]$, and define a sequence $\{Y_n\}$ on $(\Omega,\mathscr{A},P)$ by letting $\{Y_n(\omega)\}$ be the sequence of 0's and 1's in the binary expansion of $X(\omega)$. Then $\{Y_n\}$ is a sequence of independent random variables, each of which has a Bernoulli distribution with parameter $1/2$.*

(b) *Conversely, suppose that $\{Y_n\}$ is a sequence of independent random variables on $(\Omega,\mathscr{A},P)$, each of which has a Bernoulli distribution with parameter $1/2$. Then the random variable X defined by $X=\sum_n Y_n/2^n$ is uniformly distributed on the interval $[0,1]$.*

Proof. A proof for part (a) was given just before the statement of the proposition. We turn to part (b). By modifying the variables Y_n on a null set if necessary, we can assume that for every ω the sequence $\{Y_n(\omega)\}$ contains only 0's and 1's and does not end with an infinite string of 1's. Consider the dyadic rational $i/2^n$, where i satisfies $0\le i<2^n$. Then $i/2^n$ has an n-bit binary expansion, say $0.b_1b_2\ldots b_n$, and $X(\omega)$ belongs to the interval $[i/2^n,(i+1)/2^n)$ if and only if $Y_j(\omega)=b_j$ holds for $j=1,\ldots,n$. Thus $P_X(I)=\lambda(I)$ holds for intervals I of the form $[i/2^n,(i+1)/2^n)$ and hence (see Lemma 1.4.2) for all open subsets I of $(0,1)$. In view of the regularity of P_X and λ (Proposition 1.5.6), the proof is complete. □

Corollary 10.1.14. *There is an infinite sequence of independent random variables, each of which is uniformly distributed on $[0,1]$. Such a sequence can be constructed on the probability space $([0,1],\mathscr{B}([0,1]),\lambda)$.*

Proof. Let X be a random variable that is uniformly distributed on $[0,1]$; such a random variable can of course be defined on the probability space $([0,1],\mathscr{B}([0,1]),\lambda)$. Let $\{Y_n\}$ be the sequence of random variables constructed in part (a) of Proposition 10.1.13. Since the set $\mathbb{N}$ of positive integers has the same cardinality as the set $\mathbb{N}\times\mathbb{N}$ of pairs of positive integers, we can reindex the sequence $\{Y_n\}$, obtaining a doubly indexed sequence $\{Y'_{m,n}\}$. For each n define a random variable Z_n by $Z_n=\sum_m Y'_{m,n}/2^m$. Then part (b) Proposition 10.1.13 implies that each Z_n is uniformly distributed on $[0,1]$, while Proposition 10.1.7 implies that the variables $\{Z_n\}$ are independent. □

It is possible to use uniformly distributed random variables to construct random variables having arbitrary distributions on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. This can be done as follows:

Proposition 10.1.15. *Let μ be a probability measure on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ with cumulative distribution function F, and let X be a random variable that is uniformly distributed on the interval $(0,1)$. Then the function $F^{-1}\colon(0,1)\to\mathbb{R}$ defined by*

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : t \leq F(x)\}$$

is Borel measurable, and $F^{-1} \circ X$ has distribution μ.

Proof. The function F satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$, from which it follows that for each t in $(0,1)$ the set $\{x \in \mathbb{R} : t \leq F(x)\}$ is nonempty and bounded below and hence that each $F^{-1}(t)$ is finite. If $t_1 < t_2$, then

$$\{x \in \mathbb{R} : t_2 \leq F(x)\} \subseteq \{x \in \mathbb{R} : t_1 \leq F(x)\},$$

and taking the infima of these sets gives $F^{-1}(t_1) \leq F^{-1}(t_2)$. In other words, F^{-1} is nondecreasing, and so it is Borel measurable.

Let us check that

$$F^{-1}(t) \leq x \tag{3}$$

holds if and only if

$$t \leq F(x). \tag{4}$$

It is immediate from the definition of F^{-1} that (4) implies (3). On the other hand, if (3) holds, then there is a sequence $\{x_n\}$ that decreases to x and is such that $t \leq F(x_n)$ holds for each n. Since F is right continuous, (4) follows and the proof of the equivalence of (3) and (4) is complete.

Finally, the equivalence of (3) and (4) implies that for each x in $\mathbb{R}$ we have

$$P(F^{-1} \circ X \leq x) = P(X \leq F(x)) = F(x);$$

thus $F^{-1} \circ X$ has distribution function F and distribution μ. □

Corollary 10.1.16. *Let μ be a probability distribution on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Then there is an infinite sequence of independent random variables, each of which has distribution μ. Such a sequence of random variables can be constructed on the probability space $([0,1], \mathscr{B}([0,1]), \lambda)$.*

Proof. This is an immediate consequence of Corollary 10.1.14 and Proposition 10.1.15. □

Given a source of independent and uniformly distributed random numbers (for instance, a table of random numbers or a random number generator on a computer), one can use the techniques of Proposition 10.1.15 and Corollary 10.1.16 to simulate a sequence of observations from an arbitrary distribution.

Exercises

1. Let $(\Omega, \mathscr{A}, P)$ be a probability space, and let $A_1, A_2, \ldots, A_n$ be a finite indexed family of events in $\mathscr{A}$. Show that the conditions

(i) the events $A_1, A_2, \ldots, A_n$ are independent,
(ii) the equation

$$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n)$$

holds for every choice of $B_1, B_2, \ldots, B_n$, where for each i the event B_i is either A_i or A_i^c,
(iii) the events $A_1^c, A_2^c, \ldots, A_n^c$ are independent, and
(iv) the random variables $\chi_{A_1}, \chi_{A_2}, \ldots, \chi_{A_n}$ are independent

are equivalent.

2. Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $X_1, \ldots, X_d$ be real-valued random variables on Ω, and let X be the $\mathbb{R}^d$-valued random vector whose components are $X_1, \ldots, X_d$. Suppose that $F_{X_1}, \ldots, F_{X_d}$ are the cumulative distribution functions of $X_1, \ldots, X_d$ and that F_X is the cumulative distribution function of X, defined by

$$F_X(t_1, \ldots, t_d) = P(X_i \leq t_i \text{ for all } i).$$

Show that $X_1, \ldots, X_d$ are independent if and only if

$$F_X(t_1, \ldots, t_d) = F_{X_1}(t_1)\cdots F_{X_d}(t_d)$$

holds for all $(t_1, \ldots, t_d)$ in $\mathbb{R}^d$. (Hint: Use Theorem 1.6.2.)

3. Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $X_1, \ldots, X_d$ be real-valued random variables on Ω, and let X be the $\mathbb{R}^d$-valued random vector whose components are $X_1, \ldots, X_d$. Let $\mu_1, \ldots, \mu_d$ be the distributions of $X_1, \ldots, X_d$, and let μ be the distribution of X.

(a) Show that if μ is absolutely continuous (with respect to Lebesgue measure), then $\mu_1, \ldots, \mu_d$ are absolutely continuous.
(b) Show by example that the absolute continuity of μ does not follow from the absolute continuity of $\mu_1, \ldots, \mu_d$.

4. Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $X_1, \ldots, X_d$ be real-valued random variables on Ω, and let X be the $\mathbb{R}^d$-valued random vector whose components are $X_1, \ldots, X_d$. Suppose that the distributions of $X_1, \ldots, X_d$ are absolutely continuous, with densities $f_1, \ldots, f_d$. Show that $X_1, \ldots, X_d$ are independent if and only if the random vector X is an absolutely continuous random variable whose density is given by $(t_1, \ldots, t_d) \mapsto f_1(t_1)\ldots f_d(t_d)$.

5. Let $X_1, X_2, \ldots, X_n$ be independent random variables, each of which has a Bernoulli distribution with parameter p, and let $S = X_1 + X_2 + \cdots + X_n$.

(a) Show that S has a *binomial distribution with parameters n and p*—that is, that it is concentrated on the set $\{0, 1, \ldots, n\}$, with $P(S = k)$ being given by $\binom{n}{k} p^k (1-p)^{n-k}$ for each k in $\{0, 1, \ldots, n\}$.
(b) Show that $E(S) = np$ and $\text{var}(S) = np(1-p)$.

6. A real-valued random variable has a *Poisson distribution with parameter* λ if its values are nonnegative integers, with $P(X=k)=\lambda^k e^{-\lambda}/k!$ for each nonnegative integer k.
   (a) Check that the formula above indeed defines a probability measure on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$.
   (b) Verify that if the random variable X has a Poisson distribution with parameter λ, then $E(X)=\lambda$ and $\mathrm{var}(X)=\lambda$.
   (c) Show that if X_1 and X_2 are independent random variables that have Poisson distributions with parameters λ_1 and λ_2, respectively, then X_1+X_2 has a Poisson distribution with parameter $\lambda_1+\lambda_2$.
7. Let X_1 and X_2 be independent random variables, each of which is uniformly distributed on the interval $[0,1]$. Find the density function of X_1+X_2.
8. Let X and Y be independent normal random variables with mean 0 and variance 1, and let R and Θ be random variables with values in $[0,+\infty)$ and $[0,2\pi)$ that correspond to writing (X,Y) in polar coordinates.
   (a) Show that R and Θ are independent, that R has distribution function given by $t\mapsto 1-e^{-t^2/2}$ for nonnegative t, and that Θ has a uniform distribution.
   (b) Derive from this a way to use Proposition 10.1.15 to simulate values for normally distributed random variables by using easily available functions, rather than by using the inverse of the distribution function of a normal distribution.

10.2 Laws of Large Numbers

This section contains an introduction to the laws of large numbers.

Let X and X_1, X_2, ... be random variables on the probability space $(\Omega,\mathscr{A},P)$. Then $\{X_n\}$ is said to *converge in probability* to X if

$$\lim_n P(|X_n-X|>\varepsilon)=0$$

holds for each positive number ε and to *converge almost surely* to X (or to *converge a.s.* to X) if

$$P(X=\lim_n X_n)=1.$$

In other words, $\{X_n\}$ converges to X in probability if it converges to X in measure, and $\{X_n\}$ converges to X almost surely if it converges to X almost everywhere.[5] Thus a number of relationships between convergence in probability and almost sure convergence can be found in Chap. 3.

[5]More generally, an arbitrary (probabilistic) assertion holds *almost surely* if it holds almost everywhere.

Random variables X_i, $i \in I$, are said to be *identically distributed* if they all have the same distribution—that is, if $P_{X_i} = P_{X_j}$ for all i, j in I. Sequences $\{X_n\}$ of random variables that are independent and identically distributed occur frequently, and one often abbreviates a little and calls such sequences *i.i.d.*

Theorem 10.2.1 (Weak Law of Large Numbers). *Let $\{X_n\}$ be a sequence of independent identically distributed real-valued random variables with finite second moments. For each n let $S_n = X_1 + \cdots + X_n$. Then S_n/n converges to $E(X_1)$ in probability.*

Proof. Let ε be a positive number. Since $\operatorname{var}(S_n/n) = (1/n)\operatorname{var}(X_1)$ (see Corollary 10.1.11 and Lemma 10.1.3), Proposition 2.3.10 implies that

$$P\left(\left|\frac{S_n}{n} - E(X_1)\right| > \varepsilon\right) = P\left(\left|\frac{S_n - E(S_n)}{n}\right|^2 > \varepsilon^2\right)$$
$$\leq \frac{1}{\varepsilon^2}\operatorname{var}(S_n/n) = \frac{\operatorname{var}(X_1)}{n\varepsilon^2}.$$

Thus $\lim_n P(|\frac{S_n}{n} - E(X_1)| > \varepsilon) = 0$, and so S_n/n converges to $E(X_1)$ in probability. □

Suppose that $(\Omega, \mathscr{A}, P)$ is a probability space and that $\{A_n\}$ is a sequence of events in $\mathscr{A}$. Then

$$\{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\}$$

is equal to $\cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$; it is the event that infinitely many of the events A_n occur, and it is often written as $\{A_n \text{ i.o.}\}$ ("i.o." is an abbreviation for "infinitely often"). For example, if we are dealing with an infinite sequence of tosses of a coin, and if for each n we let A_n be the event that a head appears on the nth toss, then $\{A_n \text{ i.o.}\}$ is the event that a head appears on infinitely many of the tosses.

Proposition 10.2.2 (Borel–Cantelli Lemmas). *Let $(\Omega, \mathscr{A}, P)$ be a probability space, and let $\{A_n\}$ be a sequence of events in $\mathscr{A}$.*

(a) *If $\sum_n P(A_n) < +\infty$, then $P(\{A_n \text{ i.o.}\}) = 0$.*
(b) *If the events A_n, $n = 1, 2, \ldots$, are independent and if $\sum_n P(A_n) = +\infty$, then $P(\{A_n \text{ i.o.}\}) = 1$.*

Note that part (b) of Proposition 10.2.2 implies that if the events $\{A_n\}$ are independent and satisfy $P(\{A_n \text{ i.o.}\}) = 0$, then $\sum_n P(A_n) < +\infty$. Combining this with part (a) of the proposition, we see that for independent events the conditions $P(\{A_n \text{ i.o.}\}) = 0$ and $\sum_n P(A_n) < +\infty$ are equivalent.

Proof. Since $\{A_n \text{ i.o.}\} = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$, we have

$$P(\{A_n \text{ i.o.}\}) \leq P(\cup_{n=m}^{\infty} A_n) \leq \sum_{n=m}^{\infty} P(A_n)$$

for each m. Thus if $\sum_n P(A_n) < +\infty$, then $P(\{A_n \text{ i.o.}\}) \le \lim_m \sum_{n=m}^{\infty} P(A_n) = 0$ and so $P(\{A_n \text{ i.o.}\}) = 0$; with this, part (a) is proved.

To prove part (b), let us look at the complement of $\{A_n \text{ i.o.}\}$. We have

$$\{A_n \text{ i.o.}\}^c = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n^c,$$

and so we can prove that $P(\{A_n \text{ i.o.}\}) = 1$ by checking that $P(\cap_{n=m}^{\infty} A_n^c) = 0$ holds for each m. Since the events $A_n^c, A_{n+1}^c, \dots$ are independent (see Exercise 10.1.1), we have

$$P(\cap_{n=m}^{\infty} A_n^c) = \prod_{n=m}^{\infty} (1 - P(A_n)).$$

We can now derive the relation

$$\prod_{n=m}^{\infty} (1 - P(A_n)) = 0 \tag{1}$$

from the hypothesis that $\sum_n P(A_n) = +\infty$: If $P(A_n) = 1$ for some n that is greater than or equal to m, or if there is a positive ε such that $P(A_n) \ge \varepsilon$ holds for infinitely many n, then (1) certainly holds. Otherwise, $\log(1 - P(A_n))$ is asymptotic to $-P(A_n)$, and so $\sum_{n=m}^{\infty} \log(1 - P(A_n)) = -\infty$, from which (1) follows. □

Proposition 10.2.3 (Kolmogorov's Zero–One Law). *Suppose that $\{X_n\}$ is a sequence of independent random variables. Then each event that belongs to the σ-algebra $\cap_n \sigma(X_n, X_{n+1}, \dots)$ has probability* 0 *or* 1.

The intersection of the σ-algebras $\sigma(X_n, X_{n+1}, \dots)$ is, of course, a σ-algebra. It is called the *tail σ-algebra* of the sequence $\{X_n\}$, and its members are called *tail events*. Thus Kolmogorov's zero–one law can be rephrased so as to say that each tail event of a sequence of independent random variables has probability 0 or 1.

Proof. Let $\mathcal{T}$ be the tail σ-algebra for the sequence $\{X_n\}$. Proposition 10.1.7 implies that for each n the σ-algebras $\sigma(X_1)$, …, $\sigma(X_{n-1})$, and $\sigma(X_n, X_{n+1}, \dots)$ are independent and hence that $\sigma(X_1)$, …, $\sigma(X_{n-1})$, and $\mathcal{T}$ are independent. Since this is true for every n, it follows that the collection consisting of $\sigma(X_n)$, $n = 1$, 2, …, together with $\mathcal{T}$, is independent. Applying Proposition 10.1.7 once more shows that $\sigma(X_1, X_2, \dots)$ and $\mathcal{T}$ are independent. Since $\mathcal{T}$ is a sub-σ-algebra of $\sigma(X_1, X_2, \dots)$, $\mathcal{T}$ must be independent of $\mathcal{T}$. Thus each A in $\mathcal{T}$ satisfies $P(A) = P(A \cap A) = P(A)P(A)$, from which it follows that $P(A) = 0$ or $P(A) = 1$. □

Example 10.2.4. Suppose that $\{X_n\}$ is a sequence of independent random variables, and for each n let $S_n = X_1 + \cdots + X_n$. For each k the convergence or divergence of the sequence $\{S_n(\omega)\}$ does not depend on the values $X_1(\omega), \dots, X_k(\omega)$ but only on the later terms in the sequence $\{X_n(\omega)\}$. Thus the event $\{\lim_n S_n \text{ exists}\}$ is a tail event and so by Kolmogorov's zero–one law has probability 0 or 1. A similar argument shows that the event $\{\lim_n S_n/n \text{ exists}\}$ has probability 0 or 1. □

Theorem 10.2.5 (Strong Law of Large Numbers). *Let $\{X_n\}$ be a sequence of independent identically distributed random variables with finite expected values. For each n let $S_n = X_1 + \cdots + X_n$. Then $\{S_n/n\}$ converges to $E(X_1)$ almost surely.*

We will need the following two results for the proof of the strong law of large numbers.

Proposition 10.2.6 (Kolmogorov's Inequality). *Let $X_1, X_2, \ldots, X_n$ be independent random variables, each of which has mean 0 and a finite second moment, and for each i let $S_i = X_1 + \cdots + X_i$. Then*

$$P(\max_{1\le i\le n} |S_i| > \varepsilon) \le (1/\varepsilon^2) \sum_{i=1}^{n} E(X_i^2)$$

holds for each positive ε.

Proof. Define events A and $A_1, \ldots, A_n$ by $A = \{\max_i |S_i| > \varepsilon\}$ and

$$A_i = \{|S_i| > \varepsilon \text{ and } |S_j| \le \varepsilon \text{ for } j = 1, 2, \ldots, i-1\}.$$

Let us check that for each i we have

$$\int_{A_i} S_i^2 \, dP \le \int_{A_i} S_n^2 \, dP. \tag{2}$$

To see this, note that the random variables $\chi_{A_i} S_i$ and $S_n - S_i$ are independent, while $E(S_n - S_i) = 0$, and so Proposition 10.1.10 implies that $\int_{A_i} S_i(S_n - S_i) = 0$. Hence, if we write S_n^2 as $(S_i + (S_n - S_i))^2$ and expand, we find

$$\begin{aligned}\int_{A_i} S_n^2 \, dP &= \int_{A_i} S_i^2 \, dP + 2\int_{A_i} S_i(S_n - S_i)\, dP + \int_{A_i} (S_n - S_i)^2 \, dP \\ &= \int_{A_i} S_i^2 \, dP + \int_{A_i} (S_n - S_i)^2 \, dP \\ &\ge \int_{A_i} S_i^2 \, dP,\end{aligned}$$

and (2) follows. Using Proposition 2.3.10 and relation (2), we find

$$\varepsilon^2 P(A) = \sum_i \varepsilon^2 P(A_i) \le \sum_i \int_{A_i} S_i^2 \le \sum_i \int_{A_i} S_n^2 \le \int S_n^2;$$

since the variables X_i are independent and have mean 0, we have $E(S_n^2) = \sum E(X_i^2)$, and the proof is complete. □

Proposition 10.2.7. *Let $\{X_n\}$ be a sequence of independent random variables that have mean 0 and satisfy $\sum_n E(X_n^2) < +\infty$. Then $\sum_n X_n$ converges almost surely.*

Proof. For each n define S_n by $S_n = X_1 + X_2 + \cdots + X_n$. If for each m and n such that $m > n$ we apply Kolmogorov's inequality (Proposition 10.2.6) to the sequence $X_{n+1}, \ldots, X_m$ and then let m approach infinity, we find

$$P(\{\sup_{i>n} |S_i - S_n| > \varepsilon\}) \leq \frac{1}{\varepsilon^2} \sum_{i=n+1}^{\infty} E(X_i^2).$$

Choose a sequence $\{\varepsilon_k\}$ of positive numbers that decreases to 0, and for each k choose a positive integer n_k such that $\sum_{i=n_k+1}^{\infty} E(X_i^2) < \varepsilon_k^2/2^k$. For each k define A_k by $A_k = \{\sup_{i>n_k} |S_i - S_{n_k}| > \varepsilon_k\}$. Then

$$\sum_k P(A_k) < \sum_k \frac{1}{\varepsilon_k^2} \frac{\varepsilon_k^2}{2^k} = \sum_k 1/2^k < +\infty,$$

and so $P(\{A_k \text{ i.o.}\}) = 0$. However, for each ω outside $\{A_k \text{ i.o.}\}$ the sequence $\{S_n(\omega)\}$ is a Cauchy sequence, and so $\{S_n\}$ converges almost surely. □

Proof of Strong Law of Large Numbers. For each i let Y_i be the truncated version of X_i defined by

$$Y_i(\omega) = \begin{cases} X_i(\omega) & \text{if } |X_i(\omega)| \leq i, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Of course, the variables $\{Y_i\}$ are independent and have finite expected values.

Claim. The series $\sum_i \frac{Y_i - E(Y_i)}{i}$ converges almost surely.

Since $E((Y_i - E(Y_i))^2) \leq E(Y_i^2)$, the claim will follow from Proposition 10.2.7 if we verify that $\sum_i E(Y_i^2/i^2) < +\infty$. Let μ be the common distribution of the X_i's, and for each positive integer j define I_j by $I_j = \{x \in \mathbb{R} : j - 1 < |x| \leq j\}$. There is a constant C such that $\sum_{i=j}^{\infty} 1/i^2 \leq C/j$ holds for each j (use basic calculus), and so

$$\begin{aligned} \sum_i E(Y_i^2/i^2) &= \sum_i \frac{1}{i^2} \int_{[-i,i]} x^2 \, \mu(dx) \\ &= \sum_i \sum_{j \leq i} \frac{1}{i^2} \int_{I_j} x^2 \, \mu(dx) = \sum_j \sum_{i \geq j} \frac{1}{i^2} \int_{I_j} x^2 \, \mu(dx) \\ &\leq \sum_j C \int_{I_j} \frac{x^2}{j} \, \mu(dx) \leq C \int_{\mathbb{R}} |x| \, \mu(dx) = CE(|X_1|) < +\infty. \end{aligned}$$

With this the claim is proved.

For each n let T_n be $\sum_{i=1}^{n} \frac{Y_i - E(Y_i)}{i}$, the nth partial sum of $\sum_i \frac{Y_i - E(Y_i)}{i}$. The plan is to relate the partial sums of $\sum_i (Y_i - E(Y_i))$ to the T_n's and to $\{S_n/n\}$; this will give us the information that we need about the sequence $\{S_n/n\}$. We begin by noting that

$$\sum_{i=1}^{n}(Y_i - E(Y_i)) = \sum_{i=1}^{n} i(T_i - T_{i-1})$$
$$= nT_n - \sum_{i=1}^{n-1} T_i.$$

Since (by the claim above) $\lim_n T_n$ exists almost surely, if we divide both sides of the preceding equation by n and use item B.7 in Appendix B, we find

$$\lim_n \frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i)) = \lim_n \left(T_n - \frac{1}{n}\sum_{i=1}^{n-1} T_i\right) = 0 \quad \text{a.s.} \tag{3}$$

As preparation for the final step we check that

$$\lim_n \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = 0 \quad \text{a.s.} \tag{4}$$

and that

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = E(X_1). \tag{5}$$

Let us begin with Eq. (4). Note that the finiteness of $E(|X_1|)$ and Exercise 2.4.6 imply that $\sum_i P(\{X_i \neq Y_i\}) = \sum_i P(|X_i| > i) < +\infty$; from this and the Borel–Cantelli lemma, we conclude that $P(\{X_i \neq Y_i \text{ i.o.}\}) = 0$ and hence that (4) holds. Equation (5) follows from the fact that $\lim_i E(Y_i) = E(X_1)$, plus another use of B.7. Finally, Eqs. (3) and (5) imply that

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} Y_i = E(X_1)$$

holds almost surely, and from this, together with (4), we conclude that $\lim_n S_n/n = E(X_1)$ holds almost surely. With this the proof of the strong law is complete. □

Theorem 10.2.8 (Converse to the Strong Law of Large Numbers). *Let $\{X_n\}$ be a sequence of independent identically distributed random variables that do not have finite expected values. For each n let $S_n = X_1 + \cdots + X_n$. Then $\limsup_n |S_n/n| = +\infty$ almost surely.*

Proof. Let K be a positive integer, fixed for a moment, and for each n let A_n be the event $\{|X_n| \geq Kn\}$. Since the variables $\{X_i\}$ have a common distribution, but do not have a finite expected value, it follows from Exercise 2.4.6 that $\sum_n P(A_n) = +\infty$. The second part of the Borel–Cantelli lemmas implies that $P(\{A_n \text{ i.o.}\}) = 1$ and hence that

$$P\left(\limsup_n \frac{|X_n|}{n} \geq K\right) = 1.$$

This is true for each positive integer K, and so it follows that $\limsup_n |X_n/n| = +\infty$ almost surely. However,

$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n}\frac{S_{n-1}}{n-1},$$

from which it follows that $\limsup_n |X_n/n| \le 2\limsup_n |S_n/n|$; thus $\limsup |S_n/n|$ is also almost surely infinite. □

Exercises

1. The Weierstrass approximation theorem says that every continuous function on a closed bounded subinterval of $\mathbb{R}$ can be uniformly approximated by polynomials. This exercise is devoted to a derivation of the Weierstrass approximation theorem for functions on $[0,1]$ from the weak law of large numbers.

   Let f be a continuous real-valued function on $[0,1]$, let $\{X_n\}$ be a sequence of independent random variables, each of which has a Bernoulli distribution with parameter p, and for each n let $S_n = X_1 + \cdots + X_n$ and $Y_n = S_n/n$. For each p in $[0,1]$ let $g_n(p)$ be $E_p(f \circ Y_n)$, the expected value of $f \circ Y_n$ in the case where the underlying Bernoulli distribution has parameter p. Then (see Exercise 10.1.5)

   $$g_n(p) = \sum_{k=0}^{n} f(k/n)\binom{n}{k} p^k (1-p)^{n-k},$$

   and so g_n is a polynomial in p. Show that the sequence $\{g_n\}$ converges uniformly to f. (Hint: The weak law of large numbers says that for each ε we have $\lim_n P(|S_n/n - p| > \varepsilon) = 0$; check that this convergence is uniform in p. Use this and the uniform continuity of f to conclude that the convergence of $E_p(f \circ Y_n)$ to $f(p)$ is uniform in p.)
2. Suppose that $\{X_n\}$ is a sequence of independent random variables and that $\mathscr{T}$ is the σ-algebra of tail events of $\{X_n\}$. Show that every $[-\infty,+\infty]$-valued random variable that is $\mathscr{T}$-measurable is almost surely constant.
3. Let b be an integer such that $b \ge 2$. The digits that can occur in a base b expansion of a number are, of course, 0, 1, …, $b-1$. A number x in $[0,1]$ is *normal to base* b if each value in $\{0,1,\ldots,b-1\}$ occurs the expected fraction (namely $1/b$) of the time in the base b expansion of x—that is, if

   $$\lim_n \frac{\text{number of times } k \text{ occurs among the first } n \text{ digits of } x}{n} = \frac{1}{b}$$

   holds for $k = 0, 1, \ldots, b-1$. The value x is *normal* if it is normal to base b for every b.

(a) For a given b, show that almost every number in $[0,1]$ is normal to base b. (Hint: Modify part (a) of Proposition 10.1.13 and use the strong law of large numbers.)
(b) Conclude that almost every number in $[0,1]$ is normal.

4. (The Glivenko–Cantelli Theorem) Let $(\Omega, \mathscr{A}, P)$ be a probability space, let μ be a probability distribution on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, let F be its distribution function, and let $\{X_n\}$ be a sequence of independent random variables on $(\Omega, \mathscr{A}, P)$, each of which has distribution μ. For each ω in Ω, $\{X_n(\omega)\}$ is a sequence of real numbers, and we can define a sequence $\{\mu_n^\omega\}_{n=1}^\infty$ of measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ by letting $\mu_n^\omega = (1/n)\sum_{k=1}^n \delta_{X_k(\omega)}$. Also, let F_n^ω be the distribution function of the measure μ_n^ω; thus,

$$F_n^\omega(x) = (1/n)\sum_1^n \chi_{(-\infty,x]} \circ X_k(\omega)$$
$$= \frac{\text{number of } k \text{ in } \{1,2,\ldots,n\} \text{ for which } X_k(\omega) \le x}{n}$$

holds for all n, ω, and x. (Such functions F_n^ω are called *empirical distribution functions*.) Since μ describes the distribution of values of the X_n's, it seems plausible that for a typical ω, the measures μ_n^ω might approach μ as n becomes large. This is in fact true, and the Glivenko–Cantelli theorem makes a rather strong version of this precise, namely that for all ω outside some set of probability zero, the sequence $\{F_n^\omega(x)\}_{n=1}^\infty$ converges to $F(x)$, *uniformly* in x.
(a) As a first step, show that if $x \in \mathbb{R}$, then $F(x) = \lim_n F_n^\omega(x)$ and $F(x-) = \lim_n F_n^\omega(x-)$ hold for almost every ω in Ω.
(b) Show that if ε is a positive number, if $x_1, x_2, \ldots, x_k$ are real numbers such that $x_1 < x_2 < \cdots < x_k$ and such that the intervals $(-\infty, x_1)$, (x_1, x_2), $\ldots$, $(x_k, +\infty)$ all have measure less than ε under μ, and if ω is such that $\lim_n F_n^\omega(x_i) = F(x_i)$ and $\lim_n F_n^\omega(x_i-) = F(x_i-)$ hold for $i = 1, 2, \ldots, k$, then $\sup_x |F_n^\omega(x) - F(x)| \le \varepsilon$ holds for all large n.
(c) Use parts (a) and (b) to prove the Glivenko–Cantelli theorem.

5. Let $\{X_n\}$ be a sequence of independent identically distributed random variables that are nonnegative and satisfy $E(X_n) = +\infty$ for each n. Show that $\lim_n \frac{S_n}{n} = +\infty$ almost surely.

6. (a) Let X_1, X_2, $\ldots$, X_n be independent random variables on $(\Omega, \mathscr{A}, P)$, each of which has mean 0, for each i let $S_i = X_1 + X_2 + \cdots + X_i$, let c be a positive constant such that $|X_i| \le c$ holds almost surely for each i, and for each i let σ_i^2 be the variance of X_i. Show that for each positive number a,

$$P(\max_i |S_i| > a) \ge 1 - \frac{(a+c)^2}{\sum_i \sigma_i^2}.$$

(Hint: Start by using ideas from the proof of Kolmogorov's inequality to show that

$$E(S_n^2) \le a^2(1-P(A)) + (a+c)^2 P(A) + \sum_i (\sigma_{i+1}^2 + \cdots \sigma_n^2) P(A_i),$$

where $A_1, \ldots, A_n$ are given by

$$A_i = \{|S_i| > a \text{ and } |S_j| \le a \text{ for } j = 1, 2, \ldots, i-1\}$$

and $A = \cup_i A_i$.)

(b) Let $X_1, X_2, \ldots$ be independent random variables on $(\Omega, \mathscr{A}, P)$, each of which has mean 0, and for each i let σ_i^2 be the variance of X_i. Show that if there is a constant c such that $|X_i| \le c$ holds almost surely for each i and if the series $\sum_i X_i$ is almost surely convergent, then $\sum_i \sigma_i^2 < +\infty$.

(c) Show that part (b) remains true if the assumption that each X_i has mean 0 is omitted. (Hint: Define random variables $Y_1, Y_2, \ldots$ on the product of $(\Omega, \mathscr{A}, P)$ with itself by letting $Y_i(\omega_1, \omega_2) = X_i(\omega_1) - X_i(\omega_2)$, and apply part (b) to the series $\sum_i Y_i$.)

7. Let $\{X_n\}$ be a sequence of independent random variables such that $P(X_n = 1) = P(X_n = -1) = \frac{1}{2}$ holds for each n, and let $\{a_n\}$ be a sequence of real numbers. Show that the series $\sum_n a_n X_n$ converges almost surely if and only if $\{a_n\} \in \ell^2$. (Hint: See Exercise 6.)

8. Let $X_1, X_2, \ldots$ be independent random variables on $(\Omega, \mathscr{A}, P)$, let c be a positive constant, and for each i define a new random variable, the truncation $X_i^{(c)}$ of X_i by c, as follows:

$$X_i^{(c)}(\omega) = \begin{cases} X_i(\omega) & \text{if } |X_i(\omega)| \le c, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

The *three series theorem* says that the series $\sum_i X_i$ converges almost surely if and only if the series

(i) $\sum_i P(|X_i| > c)$,
(ii) $\sum_i E(X_i^{(c)})$, and
(iii) $\sum_i \operatorname{var}(X_i^{(c)})$

all converge. Prove the three series theorem. (Hint: Use the Borel–Cantelli lemma, Proposition 10.2.7, and Exercise 6.)

10.3 Convergence in Distribution and the Central Limit Theorem

In this section we look at circumstances under which probability distributions on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, or on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, give good approximations to one another. As a

rather trivial example, if n is large, then the point mass $\delta_{1/n}$ concentrated at $1/n$ should be considered to be close to the point mass δ_0 concentrated at 0. As a somewhat less trivial example, for large values of n the measure $(1/n)\sum_{i=1}^{n}\delta_{i/n}$ would seem to give a reasonable approximation to the uniform distribution on $[0,1]$. More significantly, we will see in Theorem 10.3.16 (the central limit theorem) that the distributions of certain normalized sums of random variables are well approximated by Gaussian distributions.

We should note that for our current purposes the total variation norm (defined in Sect. 4.1) does not lead to a reasonable criterion for closeness. For example, the total variation distance between $\delta_{1/n}$ and δ_0 is 2, however large n is. We need a definition that, for large n, will classify these measures as close.

We will deal with such questions in terms of convergence of sequences of probability measures (for a bit about an approach using distances, see Exercise 12 and the notes at the end of the chapter). Let μ and $\mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. The sequence $\{\mu_n\}$ is said to *converge in distribution*, or to *converge weakly*, to μ if

$$\int f\,d\mu = \lim_n \int f\,d\mu_n$$

holds for each bounded continuous f on $\mathbb{R}^d$.

Before doing anything else, we should verify that limits in distribution of sequences of probability measures are unique. In other words, we should check that if the sequence $\{\mu_n\}$ converges in distribution to μ and to ν, then $\mu = \nu$. This, however, is an immediate consequence of the following lemma.

Lemma 10.3.1. *Let μ and ν be probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. If $\int f\,d\mu = \int f\,d\nu$ holds for each bounded continuous f on $\mathbb{R}^d$, then $\mu = \nu$.*

Lemma 10.3.1 is an immediate consequence of the Riesz representation theorem (Theorem 7.2.8). The following proof, however, does not depend on the Riesz representation theorem and so avoids unnecessary dependence on Chap. 7.

Proof. Since μ and ν are regular (see Proposition 1.5.6), it is enough to prove that each compact subset K of $\mathbb{R}^d$ satisfies $\mu(K) = \nu(K)$. So let K be a nonempty compact subset of $\mathbb{R}^d$. Recall that the distance $d(x,K)$ between the point x and the set K is continuous as a function of x (see D.27) and is equal to 0 exactly when $x \in K$. For each k define a function $f_k\colon \mathbb{R}^d \to \mathbb{R}$ by $f_k(x) = \max(0, 1 - kd(x,K))$. These functions are bounded (by 0 and 1) and continuous, and they form a sequence that decreases to the indicator function χ_K of K. Furthermore $\int f_k\,d\mu = \int f_k\,d\nu$ holds for each k, and so we can use the dominated convergence theorem (or the monotone convergence theorem) to conclude that

$$\mu(K) = \lim_k \int f_k\,d\mu = \lim_k \int f_k\,d\nu = \nu(K).$$

With this the proof of the lemma is complete. □

Proposition 10.3.2. *Suppose that μ and μ_n, $n = 1, 2, \ldots$, are probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Then the conditions*

(a) *the sequence $\{\mu_n\}$ converges in distribution to μ,*
(b) *each bounded uniformly continuous f on $\mathbb{R}^d$ satisfies $\int f\,d\mu = \lim_n \int f\,d\mu_n$,*
(c) *each closed subset F of $\mathbb{R}^d$ satisfies $\limsup_n \mu_n(F) \le \mu(F)$,*
(d) *each open subset U of $\mathbb{R}^d$ satisfies $\mu(U) \le \liminf_n \mu_n(U)$, and*
(e) *each Borel subset B of $\mathbb{R}^d$ whose boundary has measure 0 under μ satisfies $\mu(B) = \lim_n \mu_n(B)$*

are equivalent.

Proof. Since every uniformly continuous function is continuous, condition (b) is an immediate consequence of condition (a). Now assume that condition (b) holds. If F is a nonempty closed subset of $\mathbb{R}^d$, then the functions $f_k\colon \mathbb{R}^d \to \mathbb{R}$ defined by $f_k(x) = \max(0, 1 - kd(x,F))$ are bounded (by 0 and 1) and uniformly continuous (again see D.27). Since these functions decrease to the indicator function of F, it follows that $\mu(F) = \lim_k \int f_k\,d\mu$. Now suppose that ε is a positive constant, and choose k such that $\int f_k\,d\mu < \mu(F) + \varepsilon$. Then, since $\mu_n(F) \le \int f_k\,d\mu_n$ holds for each n, we have

$$\limsup_n \mu_n(F) \le \lim_n \int f_k\,d\mu_n = \int f_k\,d\mu < \mu(F) + \varepsilon,$$

and condition (c) follows. It is easy to check that condition (d) is equivalent to condition (c). Now suppose that conditions (c) and (d) hold, and let B be a Borel set whose boundary has μ-measure 0. Let F and U be the closure and interior of B. Then $F - U$ is the boundary of B, and so $\mu(F) = \mu(U) = \mu(B)$, from which it follows that

$$\begin{aligned}\mu(B) = \mu(U) &\le \liminf_n \mu_n(U)\\ &\le \liminf_n \mu_n(B) \le \limsup_n \mu_n(B)\\ &\le \limsup_n \mu_n(F) \le \mu(F) = \mu(B).\end{aligned}$$

Thus, condition (e) follows from conditions (c) and (d).

Finally, we derive condition (a) from condition (e). So suppose that condition (e) holds, and let f be a bounded continuous function on $\mathbb{R}^d$. Suppose that ε is a positive number. Let B be a positive number such that $-B \le f(x) < B$ holds for all x, and let $c_0, c_1, \ldots, c_k$ be numbers such that

$$-B = c_0 < c_1 < \cdots < c_k = B$$

(we still need to look at the details of how the c_i's are to be chosen). For $i = 1, \ldots, k$ let $C_k = \{x \in \mathbb{R}^d : c_{k-1} \le f(x) < c_k\}$. The continuity of f implies that the boundary of C_k is included in the set of points x such that $f(x)$ is equal to c_{k-1} or c_k. Since the sets $\{x \in \mathbb{R}^d : f(x) = c\}$, where c ranges over $\mathbb{R}$, are disjoint and Borel, at most a countable number of them can have positive measure under μ. It follows that we can choose our points c_i so that the boundaries of the sets C_i have μ-measure 0 and so that each interval $[c_{i-1}, c_i)$ has length less than ε. If we define g by $g = \sum_{i=1}^{k} c_i \chi_{C_i}$, then $f \le g \le f + \varepsilon$, and so, if we apply condition (e) to the sets C_i, we find

$$\limsup_n \int f\, d\mu_n \le \lim_n \int g\, d\mu_n = \int g\, d\mu \le \int f\, d\mu + \varepsilon.$$

A similar calculation shows that $\int f\, d\mu - \varepsilon \le \liminf_n \int f\, d\mu_n$. Since ε is arbitrary, condition (a) follows, and with that the proof of the proposition is complete. □

As we have seen, probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ can be identified with distribution functions. Here is a characterization of convergence in distribution on $\mathbb{R}$ in terms of distribution functions (in fact, convergence in distribution seems to have first been defined in terms of distribution functions).

Proposition 10.3.3. *Suppose that μ and μ_n, $n = 1, 2, \ldots$, are probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, with distribution functions F and F_n, $n = 1, 2, \ldots$. Then the conditions*

(a) *$\{\mu_n\}$ converges in distribution to μ,*
(b) *$F(t) = \lim_n F_n(t)$ holds at each t at which F is continuous, and*
(c) *$F(t) = \lim_n F_n(t)$ holds at each t in some dense subset of $\mathbb{R}$*

are equivalent.

Proof. It follows from Proposition 10.3.2 that condition (a) implies condition (b) and from the fact that a monotone function has at most countably many discontinuities (see Lemma 6.3.2) that condition (b) implies condition (c). To show that condition (c) implies condition (a), we will assume that condition (c) holds and prove that each open subset U of $\mathbb{R}$ satisfies $\mu(U) \le \liminf_n \mu_n(U)$ (see Proposition 10.3.2). So suppose that U is a nonempty open subset of $\mathbb{R}$. Let ε be a positive number. According to Proposition C.4, there is a sequence $\{U_i\}$ of disjoint open intervals whose union is U. We can choose an integer k such that $\mu(U) - \varepsilon < \mu(\cup_{i=1}^{k} U_i)$. Next we approximate the sets U_i, $i = 1, \ldots, k$, with subintervals C_i such that $\sum_{i=1}^{k} \mu(U_i) - \varepsilon < \sum_{i=1}^{k} \mu(C_i)$ and such that each C_i is of the form $(c_i, d_i]$, where c_i and d_i belong to the dense set given by condition (c). Then each C_i satisfies $\mu(C_i) = \lim_n \mu_n(C_i)$, and it follows that

$$\mu(U) - 2\varepsilon < \sum_i \mu(C_i) = \lim_n \sum_i \mu_n(C_i) \le \liminf_n \mu_n(U).$$

Since ε is arbitrary, we have $\mu(U) \le \liminf_n \mu_n(U)$, and the proof is complete. □

Next we introduce the Fourier transform of a probability measure. For that we need to know a bit about the integration of complex-valued functions; see Sect. 2.6. We will also be using complex-valued exponential functions; see item B.10 in Appendix B for the facts we need.

In addition, we need the following basic result:

Lemma 10.3.4. *Let z and $\{z_n\}$, $n = 1, 2, \ldots$, be complex numbers such that $z = \lim_n z_n$. Then $\lim_n (1 + z_n/n)^n = e^z$.*

Proof. Choose a positive constant M that is larger than the absolute values of z and of every z_n. For each k the term in the binomial expansion of $(1 + z_n/n)^n$ that involves the kth power of z_n is

$$\binom{n}{k} \frac{z_n^k}{n^k}.$$

As n approaches infinity, this term approaches the term $z^k/k!$ from the series expansion of e^z. Let us check that the sum of the terms of the binomial expansion of $(1 + z_n/n)^n$ approaches the sum of the terms of the series for e^z. The issue here is the interchange of sums and limits, and this interchange can be justified with the dominated convergence theorem, if we apply that theorem to integrals (i.e., sums) on the space of nonnegative integers together with counting measure and if we note that the functions involved here are dominated by the terms in the series expansion of e^M). Thus $\lim_n (1 + z_n/n)^n = e^z$, and the proof is complete. □

Now suppose that μ is a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. The *characteristic function*,[6] or *Fourier transform*, of μ is the function $\phi_\mu \colon \mathbb{R}^d \to \mathbb{C}$ defined[7] by $\phi_\mu(t) = \int e^{i(t,x)}\, \mu(dx)$. (The integrand here is bounded and measurable, and so the definition of ϕ_μ makes sense.) If X is an $\mathbb{R}^d$-valued random variable, then the characteristic function of X, written ϕ_X, is defined to be the characteristic function of the distribution P_X of X, and so $\phi_X(t) = \phi_{P_X}(t) = E(e^{i(t,X)})$.

Proposition 10.3.5. *Let μ be a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Then*

(a) $\phi_\mu(0) = 1$,
(b) $|\phi_\mu(t)| \le 1$ *holds for each t in $\mathbb{R}^d$, and*
(c) ϕ_μ *is continuous on $\mathbb{R}^d$.*

Proof. Part (a) is immediate, and part (b) follows from Proposition 2.6.7. For part (c), let t be an arbitrary element of $\mathbb{R}^d$, and suppose that $\{t_n\}$ is a sequence of elements of $\mathbb{R}^d$ such that $t = \lim_n t_n$. Then the dominated convergence theorem

[6] The phrase "characteristic function" is ambiguous; it can mean either "Fourier transform" or "indicator function" (see item A.3 in Appendix A). In this chapter we follow the usage of probabilists and use characteristic function to mean Fourier transform; in the rest of the book we use characteristic function to mean indicator function.

[7] Here (t,x) is the inner product of t and x, defined by $(t,x) = \sum_{j=1}^d t_j x_j$. In case we are dealing with measures on $\mathbb{R}$, rather than on $\mathbb{R}^d$, we write e^{itx}, rather than $e^{i(t,x)}$.

implies that

$$\lim_n \int e^{i(t_n,x)}\,\mu(dx) = \int e^{i(t,x)}\,\mu(dx)$$

and hence that $\lim_n \phi_\mu(t_n) = \phi_\mu(t)$. Since this holds for every sequence $\{t_n\}$ that converges to t, the continuity of ϕ_μ follows (see D.31 in Appendix D). □

Lemma 10.3.6. *Let X be a real-valued random variable, let a and b be real constants, and define a random variable Y by $Y = aX + b$. Then $\phi_Y(t) = e^{itb}\phi_X(at)$ holds for all real t.*

Proof. This follows from the calculation $\phi_Y(t) = E(e^{it(aX+b)}) = e^{itb}E(e^{iatX}) = e^{itb}\phi_X(at)$. □

Proposition 10.3.7. *Let μ be a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, and let n be a positive integer such that μ has a finite nth moment—that is, such that $\int |x|^n\,\mu(dx)$ is finite. Then ϕ_μ has n continuous derivatives, which are given by*

$$\phi_\mu^{(k)}(t) = i^k \int x^k e^{itx}\,\mu(dx)$$

for $k = 1, 2, \ldots, n$.

Proof. Note[8] that $|e^{iu} - 1| \le |u|$ holds for all real u and that $\lim_{u\to 0}(e^{iu} - 1)/u = i$. We will use those facts in the calculations below.

We verify the formula for $\phi_\mu^{(k)}$ by using mathematical induction. Suppose that we have already verified that $\phi_\mu^{(k)}$ has the required form (certainly $\phi_\mu^{(0)}$ is ϕ_μ and has the required form). Then

$$\begin{aligned}\frac{\phi_\mu^{(k)}(t+h) - \phi_\mu^{(k)}(t)}{h} &= i^k \int x^k \frac{e^{i(t+h)x} - e^{itx}}{h}\,\mu(dx)\\ &= i^k \int x^k e^{itx}\frac{e^{ihx} - 1}{h}\,\mu(dx).\end{aligned}$$

The integrand in the second integral above approaches $ix^{k+1}e^{itx}$ as h approaches 0, and it is dominated by $|x^{k+1}|$. It follows from the dominated convergence theorem that if $0 \le k < n$ and if $\phi_\mu^{(k)}$ has the form given in the proposition, then $\phi_\mu^{(k+1)}$ has the analogous form with k replaced by $k+1$. (Note that, as in the proof of Proposition 10.3.5, we are actually taking limits as h approaches 0 along sequences.) The continuity of $\phi_\mu^{(k+1)}$ follows from another application of the dominated convergence theorem. □

[8] A geometric justification for the inequality $|e^{iu} - 1| \le |u|$ comes from the fact that $|e^{iu} - 1|$ is the straight-line distance between the points $(\cos u, \sin u)$ and $(1, 0)$, while $|u|$ gives the length of a path that connects these points and lies on the unit circle. Alternatively, we can give this inequality and also the limit $\lim_{u\to 0}(e^{iu} - 1)/u = i$ non-geometric proofs if we rewrite the exponentials in terms of sines and cosines.

Proposition 10.3.8. *Suppose that P is the normal distribution on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with mean μ and variance σ^2. Then $\phi_P(t) = e^{it\mu} e^{-\sigma^2 t^2/2}$.*

Proof. Let us begin with the special case where P is the standard normal distribution (i.e., the normal distribution with mean 0 and variance 1). Then the Fourier transform ϕ_P of P is given by

$$\phi_P(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{ixt} e^{-x^2/2}\, dx.$$

It is easy to check that P has a finite first moment (in fact, finite moments of all orders), and so it follows from Proposition 10.3.7 that

$$\phi_P'(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} ixe^{ixt} e^{-x^2/2}\, dx.$$

If we integrate by parts (view the integrand above as the product of ie^{ixt} and the derivative of $-e^{-x^2/2}$), we find that $\phi_P'(t) = -t\phi_P(t)$. It follows that the derivative of $t \mapsto e^{t^2/2}\phi_P(t)$ is identically zero and so, since $\phi_P(0) = 1$, that $\phi_P(t) = e^{-t^2/2}$. The general case now follows from Lemma 10.3.6. □

Proposition 10.3.9. *Let ν_1 and ν_2 be probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, and let ν be their convolution. Then $\phi_\nu(t) = \phi_{\nu_1}(t)\phi_{\nu_2}(t)$ holds at each t in $\mathbb{R}^d$.*

Proof. Let X_1 and X_2 be independent random variables with distributions ν_1 and ν_2. Then $X_1 + X_2$ has distribution ν, and so Proposition 10.1.10 implies that

$$\phi_\nu(t) = E(e^{it(X_1+X_2)}) = E(e^{itX_1})E(e^{itX_2}) = \phi_{\nu_1}(t)\phi_{\nu_2}(t). \qquad \square$$

Example 10.3.10. Let us now try to invert the Fourier transform—to go from the Fourier transform of a probability measure back to the measure. We start with the Gaussian distributions and look at $t \mapsto e^{-\sigma^2 t^2/2}$, the Fourier transform of the Gaussian distribution with mean 0 and variance σ^2. If we multiply this function by e^{-ixt}, integrate, and use Proposition 10.3.8 at the last step, we find

$$\begin{aligned}\int_{\mathbb{R}} e^{-ixt} e^{-\sigma^2 t^2/2} dt &= \int_{\mathbb{R}} e^{ixt} e^{-\sigma^2 t^2/2} dt \\ &= \frac{\sqrt{2\pi}}{\sigma} \frac{1}{\sqrt{2\pi}\frac{1}{\sigma}} \int_{\mathbb{R}} e^{ixt} e^{-\frac{t^2}{2\frac{1}{\sigma^2}}} dt \\ &= \frac{\sqrt{2\pi}}{\sigma} e^{-x^2/2\sigma^2}.\end{aligned}$$

It follows that

$$\frac{1}{2\pi} \int_{\mathbb{R}} e^{-ixt} e^{-\sigma^2 t^2/2} dt = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}. \qquad \square$$

In particular, we can go from the Fourier transform ϕ of the Gaussian distribution with mean 0 and variance σ^2 back to its density, say g, by using the *Fourier inversion formula*

$$\frac{1}{2\pi}\int_{\mathbb{R}} e^{-ixt}\phi(t)dt = g(x), \tag{1}$$

which says that the *inverse Fourier transform* of ϕ is equal to g. The Fourier inversion formula works for many distributions, but not all (see Exercise 13). However, we now have enough information to prove the following uniqueness theorem.

Proposition 10.3.11. *Let μ and ν be probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Then $\mu = \nu$ if and only if $\phi_\mu = \phi_\nu$.*

Proof. The following is a proof for measures on $\mathbb{R}$, rather than on $\mathbb{R}^d$. We can convert it to a proof for measures on $\mathbb{R}^d$ by changing the constant $1/2\pi$ in the Fourier inversion formula to $1/(2\pi)^d$, replacing e^{-ixt} with $e^{-i(x,t)}$, and checking that the Fourier inversion formula works for probabilities on $\mathbb{R}^d$ that are products of d Gaussian distributions, each with mean 0 and variance σ^2.

So let us turn to the proof when $d = 1$. It is certainly true that if $\mu = \nu$, then $\phi_\mu = \phi_\nu$, and so we need only check that if $\phi_\mu = \phi_\nu$, then $\mu = \nu$. So let μ and ν be probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $\phi_\mu = \phi_\nu$. In addition, let γ_σ be the Gaussian distribution on $\mathbb{R}$ with mean 0 and variance σ^2; let ϕ_{γ_σ} and g_σ be its Fourier transform and density function. Let us calculate the inverse Fourier transform of $\phi_{\gamma_\sigma * \mu}$, or equivalently of $\phi_{\gamma_\sigma}\phi_\mu$ (Proposition 10.3.9), using the fact that we know from Example 10.3.10 that the Fourier inversion formula works in the Gaussian case:

$$\begin{aligned}
\frac{1}{2\pi}\int_{\mathbb{R}} e^{-ixt}\phi_{\gamma_\sigma}(t)\phi_\mu(t)\,dt &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ixt}\phi_{\gamma_\sigma}(t)\int_{\mathbb{R}} e^{its}\,\mu(ds)\,dt \\
&= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-it(x-s)}\phi_{\gamma_\sigma}(t)\,dt\,\mu(ds) \\
&= \int_{\mathbb{R}} g_\sigma(x-s)\,\mu(ds)
\end{aligned}$$

(we were able to apply Fubini's theorem because μ is finite and ϕ_{γ_σ} is integrable with respect to Lebesgue measure). Note that the result of this calculation is the density of $\gamma_\sigma * \mu$ (see Proposition 10.1.12). In other words, the inverse Fourier transform of $\phi_{\gamma_\sigma}\phi_\mu$ is the density of $\gamma_\sigma * \mu$. A similar calculation can be applied to ν. Since μ and ν are such that $\phi_\mu = \phi_\nu$, we can conclude from these calculations that $\gamma_\sigma * \mu = \gamma_\sigma * \nu$. Finally, $\gamma_\sigma * \mu$ and $\gamma_\sigma * \nu$ converge in distribution to μ and ν as σ approaches 0 (check this; you might use Exercise 7), and it follows that $\mu = \nu$. □

Corollary 10.3.12. *Let $X_1, \ldots, X_d$ be real random variables, all defined on the same probability space, and let X be the $\mathbb{R}^d$-valued random variable whose*

components are $X_1, \ldots, X_d$. Then the random variables $X_1, \ldots, X_d$ are independent if and only if $\phi_X(t) = \prod_k \phi_{X_k}(t_k)$ holds for each vector $t = (t_1, \ldots, t_d)$ in $\mathbb{R}^d$.

Proof. If the random variables $X_1, \ldots, X_d$ are independent, then the relation $\phi_X(t) = \prod_k \phi_{X_k}(t_k)$ follows from Proposition 10.1.10, which can easily be extended to apply to complex-valued functions.

We turn to the converse. Let μ_X and $\mu_{X_1}, \ldots, \mu_{X_d}$ be the distributions of X and $X_1, \ldots, X_d$. Since the characteristic function (call it ϕ_{prod}) of the product measure $\mu_{X_1} \times \cdots \times \mu_{X_d}$ is given by $\phi_{\text{prod}}(t) = \prod_k \phi_{X_k}(t_k)$, it follows from Proposition 10.3.11 that the relation $\phi_X(t) = \prod_k \phi_{X_k}(t_k)$ implies that μ is equal to the product measure $\mu_{X_1} \times \cdots \times \mu_{X_d}$ and then from Proposition 10.1.9 that the random variables $X_1, \ldots, X_d$ are independent. □

Our goal for the rest of this section is to prove the central limit theorem (Theorem 10.3.16). The main tool for this will be Proposition 10.3.15.

Suppose that $\{\mu_n\}$ is a sequence of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Let us look at the relationship between convergence in distribution of the sequence $\{\mu_n\}$ and pointwise convergence of the corresponding sequence $\{\phi_{\mu_n}\}$ of characteristic functions. For this we need a concept related to regularity. We know (see Proposition 1.5.6) that if μ is a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, then

$$\sup\{\mu(K) : K \text{ is compact}\} = 1.$$

Measures satisfying this condition are sometimes called *tight*. A collection $\mathscr{C}$ of probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ is called *uniformly tight* if for every positive ε there is a compact set K such that

$$\mu(K) > 1 - \varepsilon$$

holds for each μ in $\mathscr{C}$.

The following result is sometimes useful for establishing the uniform tightness of a family of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. See, for example, the proof of Proposition 10.3.15.

Proposition 10.3.13. *Suppose that μ is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and that ϕ is its characteristic function. Then for each positive ε we have*

$$\mu\left(\left\{x \in \mathbb{R} : |x| \geq \frac{2}{\varepsilon}\right\}\right) \leq \frac{1}{\varepsilon} \int_{-\varepsilon}^{\varepsilon} (1 - \phi(t))\, dt.$$

Since characteristic functions are complex-valued functions, it's conceivable that the integral on the right-hand side of the inequality above could have a non-real value, in which case the inequality would be meaningless. We'll see in the proof below that this difficulty does not occur.

Proof. Using Fubini's theorem and basic calculus, we find

$$\begin{aligned}\int_{-\varepsilon}^{\varepsilon} \phi(t)\,dt &= \int_{-\varepsilon}^{\varepsilon} \int_{\mathbb{R}} e^{itx}\,\mu(dx)\,dt \\ &= \int_{\mathbb{R}} \int_{-\varepsilon}^{\varepsilon} (\cos tx + i \sin tx)\,dt\,\mu(dx) = \int_{\mathbb{R}} \frac{2\sin \varepsilon x}{x}\,\mu(dx).\end{aligned}$$

Since $(1 - \frac{\sin \varepsilon x}{\varepsilon x}) \geq \frac{1}{2}$ if $|\varepsilon x| \geq 2$, we have

$$\frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} (1 - \phi(t))\,dt = \int_{\mathbb{R}} \left(1 - \frac{\sin \varepsilon x}{\varepsilon x}\right) \mu(dx) \geq \frac{1}{2}\mu\left(\left\{x \in \mathbb{R} : |x| \geq \frac{2}{\varepsilon}\right\}\right)$$

and the proposition follows. □

Proposition 10.3.14. *Let $\{\mu_n\}$ be a uniformly tight sequence of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Then $\{\mu_n\}$ has a subsequence that converges in distribution to some probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.*

Proof. Suppose that $\{F_n\}$ is the sequence of distribution functions corresponding to $\{\mu_n\}$ and that $\{x_k\}$ is an enumeration of some countable dense subset D of $\mathbb{R}$. We will use a diagonal argument to choose a convergent subsequence of $\{\mu_n\}$. To begin, choose a subsequence $\{F_{1,n}\}_n$ of $\{F_n\}_n$ such that $\{F_{1,n}(x_1)\}_n$ is convergent, and then continue inductively, for each k choosing a subsequence $\{F_{k+1,n}\}_n$ of $\{F_{k,n}\}_n$ such that $\{F_{k+1,n}(x_{k+1})\}_n$ is convergent. Now take the diagonal subsequence $\{F_{j,j}\}$ of $\{F_n\}$, and let $\{\mu_{n_j}\}$ be the corresponding subsequence of $\{\mu_n\}$. We will show that $\{\mu_{n_j}\}$ converges in distribution to some probability measure μ.

We can define a function G_0 on the countable dense set D by letting $G_0(x) = \lim_j F_{j,j}(x)$ hold for each x in D. Then G_0 is a nondecreasing function and, since the sequence $\{\mu_n\}$ is uniformly tight, G_0 satisfies $\lim_{x\to-\infty} G_0(x) = 0$ and $\lim_{x\to+\infty} G_0(x) = 1$. Next, define $G\colon \mathbb{R} \to \mathbb{R}$ by

$$G(x) = \inf\{G_0(t) : t \in D \text{ and } t > x\}.$$

Then G is nondecreasing, it has limits of 0 and 1 at $-\infty$ and $+\infty$, and it is right continuous; let μ be the corresponding probability measure (recall Proposition 1.3.10). We show that the sequence $\{\mu_{n_j}\}$ converges in distribution to μ by checking that $G(x) = \lim_j F_{j,j}(x)$ holds at each x at which G is continuous. To do this, suppose that G is continuous at x, let ε be a positive number, and choose values t_0 and t_1 in D such that $t_0 < x < t_1$, $G(x) - \varepsilon < G_0(t_0)$, and $G_0(t_1) < G(x) + \varepsilon$. Note that if j is large enough that $|F_{j,j}(t_1) - G_0(t_1)| < \varepsilon$, then

$$F_{j,j}(x) \leq F_{j,j}(t_1) < G_0(t_1) + \varepsilon < G(x) + 2\varepsilon.$$

A similar calculation gives a lower bound of $G(x) - 2\varepsilon$ for $F_{j,j}(x)$, and so we can conclude that $|G(x) - F_{j,j}(x)| < 2\varepsilon$ holds for all large j. Thus $G(x) = \lim_j F_{j,j}(x)$, and Proposition 10.3.3 implies that $\{\mu_{n_j}\}$ converges in distribution to μ. □

Proposition 10.3.15. *Let μ and $\mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Then the sequence $\{\mu_n\}$ converges in distribution to μ if and only if the sequence $\{\phi_{\mu_n}\}$ converges pointwise to ϕ_μ.*

Proof. For each t the function $x \mapsto e^{itx}$ is bounded and continuous. Thus if $\{\mu_n\}$ converges in distribution to μ, then $\int e^{itx}\,\mu(dx) = \lim_n \int e^{itx}\,\mu_n(dx)$ holds for each t, and $\{\phi_{\mu_n}\}$ converges pointwise to ϕ_μ.

Let us turn to the converse and assume that $\{\phi_{\mu_n}\}$ converges pointwise to ϕ_μ. We begin by showing that the sequence $\{\mu_n\}$ is uniformly tight. Choose a positive number ε, and then use the continuity of ϕ_μ at 0 (and the fact that $\phi_\mu(0) = 1$) to choose δ such that $\frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \phi_\mu(t))\,dt < \varepsilon$. Since $\{\phi_{\mu_n}\}$ converges pointwise to ϕ_μ, we can use the dominated convergence theorem to conclude that

$$\frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \phi_{\mu_n}(t))\,dt < \varepsilon$$

holds for all large n. Proposition 10.3.13 now implies that

$$\mu_n\left(\left[-\frac{2}{\delta}, \frac{2}{\delta}\right]\right) > 1 - \varepsilon \tag{2}$$

holds for all large n. By making δ smaller, if necessary, we can make (2) hold for all n. It follows that the sequence $\{\mu_n\}$ is uniformly tight.

We now check that $\{\mu_n\}$ converges in distribution to μ. Suppose it did not. Then there would be a bounded continuous function f on $\mathbb{R}$ such that $\{\int f\,d\mu_n\}$ does not converge to $\int f\,d\mu$. Choose a subsequence $\{\mu_{n_k}\}$ of $\{\mu_n\}$ such that $\{\int f\,d\mu_{n_k}\}$ converges to a value other that $\int f\,d\mu$. The uniform tightness of $\{\mu_n\}$, which we verified above, together with Proposition 10.3.14, lets us replace $\{\mu_{n_k}\}$ with a subsubsequence that converges to some probability measure ν. Then $\nu \neq \mu$, since $\int f\,d\nu \neq \int f\,d\mu$, yet $\phi_\nu = \phi_\mu$, since $\{\phi_{\mu_{n_k}}\}$ converges to both ϕ_ν and ϕ_μ. This is impossible, and so our hypothesis that $\{\mu_n\}$ does not converge to μ must be false. □

Let us make a last preparation for the proof of the central limit theorem. Suppose that X is a random variable with mean 0 and variance 1 and that ϕ is its characteristic function. Then $\phi(0) = 1$, $\phi'(0) = 0$, $\phi''(0) = -1$, and ϕ has at least two continuous derivatives (see Proposition 10.3.7). According to l'Hospital's rule, plus the facts in the previous sentence, we have

$$\lim_{x\to 0}\frac{\phi(x) - (1 - x^2/2)}{x^2} = 0$$

and so ϕ can be written in terms of its second-degree Maclaurin polynomial $1-x^2/2$ as $\phi(x) = 1 - x^2/2 + R(x)$, where $\lim_{x\to 0} R(x)/x^2 = 0$.

Theorem 10.3.16 (Central Limit Theorem). *Let X_1, X_2, ... be a sequence of independent identically distributed random variables, with common mean μ and variance σ^2, and for each n let $S_n = X_1 + \cdots + X_n$. Then the normalized sequence $\{(S_n - n\mu)/\sigma\sqrt{n}\}$ converges in distribution to a normal (i.e., Gaussian) distribution with mean 0 and variance 1.*

Proof. Each random variable $(X_i - \mu)/\sigma$ has mean 0 and variance 1 and hence has a characteristic function ϕ that is as described just before the statement of the theorem. Since the X_i's are identically distributed, the function ϕ does not depend on the index i. Note that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}. \tag{3}$$

If we use Eq. (3), the independence of the X_i's, Lemma 10.3.6, Proposition 10.3.9, and the fact that $\lim_{x\to 0} R(x)/x^2 = 0$ (where $R(x)$ is the remainder defined just before the statement of the theorem), we find that the characteristic function of $(S_n - n\mu)/\sigma\sqrt{n}$ is given by

$$\phi\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2}{2n} + R(t/\sqrt{n})\right)^n = \left(1 - \frac{t^2/2 + \varepsilon_n}{n}\right)^n,$$

where $\varepsilon_n = -nR(t/\sqrt{n})$ and hence where $\lim_n \varepsilon_n = 0$. It follows (Lemma 10.3.4) that the characteristic functions of the normalized sums $(S_n - n\mu)/\sigma\sqrt{n}$ approach the function $t \mapsto e^{-t^2/2}$; since the limit is the characteristic function of the normal distribution with mean 0 and variance 1, the theorem follows (see Proposition 10.3.15). □

Exercises

1. For each positive integer n define a probability measure μ_n on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ by $\mu_n = (1/n)\sum_{i=1}^{n} \delta_{i/n}$. Show that the sequence $\{\mu_n\}$ converges in distribution to the uniform distribution on $[0, 1]$.
2. Suppose that μ and $\mu_1, \mu_2, \ldots$, are probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, each of which is concentrated on the integers. Show that the sequence $\{\mu_n\}$ converges in distribution to μ if and only if $\mu(\{k\}) = \lim_n \mu_n(\{k\})$ holds for each k in $\mathbb{Z}$.
3. Show that
   (a) if μ is the point mass at a, then ϕ_μ is given by $\phi_\mu(t) = e^{iat}$,
   (b) if μ is the binomial distribution with parameters n and p, then ϕ_μ is given by $\phi_\mu(t) = (1 - p(1 - e^{it}))^n$,

(c) if μ is the Poisson distribution with parameter λ, then ϕ_μ is given by $\phi_\mu(t) = e^{-\lambda(1-e^{it})}$, and
(d) if μ is the uniform distribution on the interval $[a,b]$, then ϕ_μ is given by $\phi_\mu(t) = \frac{e^{itb}-e^{ita}}{it(b-a)}$.

4. Show that if ϕ is the characteristic function of a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, then $\phi(-t) = \overline{\phi(t)}$.
5. Show that a probability measure μ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is symmetric (i.e., $\mu(-A) = \mu(A)$ holds for each A in $\mathscr{B}(\mathbb{R})$) if and only if ϕ_μ is real-valued.
6. Show that if ϕ is the characteristic function of a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, then ϕ is uniformly continuous on $\mathbb{R}$.
7. Suppose that X and $X_1, X_2, \ldots$ are real-valued random variables and that μ and $\mu_1, \mu_2, \ldots$ are their distributions. Show that if $\{X_n\}$ converges in probability to X, then $\{\mu_n\}$ converges in distribution to μ.
8. Let μ be a probability distribution on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Show that $|\phi_\mu(t)| = 1$ for some nonzero number t if and only if there exist real numbers a and b such that μ is concentrated on the set $\{a + bn : n \in \mathbb{Z}\}$. (Such a distribution is called a *lattice distribution*.)
9. Show directly (i.e., using only the definition of convergence in distribution) that if a sequence $\{\mu_n\}$ of probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ converges in distribution to some probability measure, then the sequence $\{\mu_n\}$ is uniformly tight.
10. Suppose that $\{\mu_n\}$ is a sequence of probability distributions on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ whose characteristic functions $\{\phi_{\mu_n}\}$ converge pointwise to some function $\phi\colon \mathbb{R} \to \mathbb{C}$. Show that if ϕ is continuous at 0, then there is a probability distribution μ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $\{\mu_n\}$ converges to μ in distribution.
11. For each n let μ_n be a binomial distribution with parameters n and p_n. Show that if $\{np_n\}$ is convergent, with $\lambda = \lim_n np_n$, then the sequence $\{\mu_n\}$ converges in distribution to the Poisson distribution with parameter λ. Do this
(a) by making a direct calculation of probabilities (see Exercise 2), and
(b) by using characteristic functions.
12. Suppose that for probability measures μ and ν on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ we define $d(\mu,\nu)$ by

$$d(\mu,\nu) = \inf\{\varepsilon > 0 : F_\mu(t) \le F_\nu(t+\varepsilon) + \varepsilon \text{ and } F_\nu(t) \le F_\mu(t+\varepsilon) + \varepsilon \text{ for all } t \text{ in } \mathbb{R}\}.$$

(The function d is known as *Lévy's metric*.)
(a) Show that d is a metric on the set of all probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.
(b) Suppose that μ and $\mu_1, \mu_2, \ldots$ are probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Show that the sequence $\{\mu_n\}$ converges in distribution to μ if and only if $\lim_n d(\mu_n, \mu) = 0$.

13. Suppose that μ is a probability distribution on $\mathbb{R}$ such that ϕ_μ is integrable. (Note that for the inversion formula (1) to make sense with the integral interpreted as a Lebesgue integral, ϕ_μ must be integrable.)
    (a) Show that if μ is absolutely continuous with density function g and if the inversion formula (1) is valid for ϕ_μ and g, then g is bounded and continuous.
    (b) Show that if ϕ_μ is integrable, then μ is absolutely continuous and formula (1) works. (Hint: Use some ideas and calculations from Proposition 10.3.11. In particular, consider $\int_{\mathbb{R}} h(x)p(x)\,dx$, where h ranges over the continuous functions with compact support on $\mathbb{R}$ and p is the inverse Fourier transform of $\phi_{\gamma_\sigma * \mu}$.)
14. Show how to prove the central limit theorem without using Proposition 10.3.13. (Hint: For each n let μ_n be the distribution of $(S_n - n\mu)/\sigma\sqrt{n}$. Use Markov's inequality (that is, Proposition 2.3.10), rather than Proposition 10.3.13, to show that the sequence $\{\mu_n\}$ is tight.)
15. Let μ and $\mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that the sequence $\{\mu_n\}$ converges in distribution to μ.
    (a) Suppose that X and $X_1, X_2, \ldots$ are random variables, all defined on the same probability space, whose distributions are μ and $\mu_1, \mu_2, \ldots$. Show (by giving a simple example) that it does not follow that $\{X_n\}$ converges almost surely to X.
    (b) On the other hand, show that there are random variables X and $X_1, X_2, \ldots$, all defined on the same probability space and with distributions μ and $\mu_1, \mu_2, \ldots$, such that $\{X_n\}$ converges to X almost surely. (Hint: Let F and $F_1, F_2, \ldots$ be the distribution functions of μ and $\mu_1, \mu_2, \ldots$. Then the random variables F^{-1} and $F_1^{-1}, F_2^{-1}, \ldots$ constructed from F and $F_1, F_2, \ldots$ as in Proposition 10.1.15 do what is required. To verify the almost sure convergence, use the equivalence of inequalities (3) and (4) from Sect. 10.1 to verify that $\lim_n F_n^{-1}(t) = F^{-1}(t)$ holds at each t at which F^{-1} is continuous.)

10.4 Conditional Distributions and Martingales

Suppose that $(\Omega, \mathscr{A}, P)$ is a probability space, that A and B are events in $\mathscr{A}$, and that $P(B) \neq 0$. In elementary treatments of probability, the *conditional probability of A, given B*, written $P(A|B)$, is defined by

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

Example 10.4.1. Suppose that we select a number at random from the set $\{1,2,3,4,5,6\}$, with each number in that set having probability $1/6$ of being selected. Consider events E and F, where E is the event that the number selected is even and F is the event that the number selected is not equal to 6. Then we have

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{2/6}{5/6} = 2/5$$

and

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{2/6}{3/6} = 2/3,$$

which should agree with one's intuition. □

Let us deal for a moment with a probability space $(\Omega, \mathscr{A}, P)$ such that Ω is finite and $\mathscr{A}$ contains all the subsets of Ω. Let X and Y be real-valued random variables on $(\Omega, \mathscr{A}, P)$ with values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, and let us assume that $P(Y = y_j) \neq 0$ for each j. Then $E(X|Y = y_j)$, the *conditional expectation of X, given that* $Y = y_j$, is defined by

$$E(X|Y = y_j) = \sum_i x_i P(X = x_i | Y = y_j).$$

It follows that

$$E(X|Y = y_j) = \frac{\sum x_i P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)} = \frac{\int_{Y=y_j} X \, dP}{P(Y = y_j)}. \tag{1}$$

Of course, this defines a function $y_j \mapsto E(X|Y = y_j)$ on the set of values of Y. It is convenient to have a slightly different form of the conditional expectation, with the new form being defined on the probability space $(\Omega, \mathscr{A}, P)$. Let us define $E(X|Y)\colon \Omega \to \mathbb{R}$ by letting $E(X|Y)(\omega)$ be $E(X|Y = y_j)$ for those ω that satisfy $Y(\omega) = y_j$. In other words, $E(X|Y)$ is the composition of the functions $\omega \mapsto Y(\omega)$ and $y \mapsto E(X|Y = y)$. It follows from (1) that

$$\int_B E(X|Y)\, dP = \int_B X \, dP \tag{2}$$

holds for each B of the form $\{Y = y_j\}$. Since each B in the σ-algebra $\sigma(Y)$ generated by Y is a finite disjoint union of sets of the form $\{Y = y_j\}$, it follows that (2) holds for each B in $\sigma(Y)$. Furthermore, $E(X|Y)$ is $\sigma(Y)$-measurable (in this simple example, where Ω is finite, this just means that $E(X|Y)$ is constant on each set of the form $\{Y = y_j\}$).

We are now ready to look at how these ideas generalize to arbitrary probability spaces.

Let $(\Omega, \mathscr{A}, P)$ be a probability space and let $\mathscr{B}$ be a sub-σ-algebra of $\mathscr{A}$. Suppose that X is a real-valued random variable on $(\Omega, \mathscr{A}, P)$ that has a finite expected value. A *conditional expectation of X given* $\mathscr{B}$ is a random variable Y that is $\mathscr{B}$-measurable, is integrable (that is, has a finite expected value), and satisfies

$$\int_B Y \, dP = \int_B X \, dP$$

for each B in $\mathscr{B}$. One generally writes $E(X|\mathscr{B})$ for a conditional expectation of X given $\mathscr{B}$. When one needs to be more precise, one sometimes calls an integrable $\mathscr{B}$-measurable function Y that satisfies $\int_B Y\,dP = \int_B X\,dP$ for all B in $\mathscr{B}$ a *version* of the conditional expectation of X given $\mathscr{B}$ or a version of $E(X|\mathscr{B})$.

Proposition 10.4.2. *Let $(\Omega, \mathscr{A}, P)$ be a probability space, let X be a random variable on $(\Omega, \mathscr{A}, P)$ that has a finite expected value, and let $\mathscr{B}$ be a sub-σ-algebra of $\mathscr{A}$. Then*

(a) *X has a conditional expectation given $\mathscr{B}$, and*
(b) *the conditional expectation of X given $\mathscr{B}$ is unique, in the sense that if Y_1 and Y_2 are versions of $E(X|\mathscr{B})$, then $Y_1 = Y_2$ almost surely.*

Proof. The formula $\mu(B) = \int_B X\,dP$ defines a finite signed measure on $(\Omega, \mathscr{B})$; it is absolutely continuous with respect to the restriction of P to $\mathscr{B}$. Thus the Radon–Nikodym theorem (Theorem 4.2.4), applied to μ and the restriction of P to $\mathscr{B}$, gives a $\mathscr{B}$-measurable random variable Y such that

$$\int_B Y\,dP = \mu(B) = \int_B X\,dP$$

holds for each B in $\mathscr{B}$. Thus Y is a conditional expectation of X given B. The uniqueness assertion in the Radon–Nikodym theorem gives the uniqueness of the conditional expectation. □

Proposition 10.4.3. *Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\mathscr{B}$ and $\mathscr{B}_0$ be sub-σ-algebras of $\mathscr{A}$, and let X and Y be random variables on $(\Omega, \mathscr{A}, P)$ that have finite expected values. Then*

(a) *if a and b are constants, then $E(aX + bY|\mathscr{B}) = aE(X|\mathscr{B}) + bE(Y|\mathscr{B})$ almost surely,*[9]
(b) *if $X \le Y$, then $E(X|\mathscr{B}) \le E(Y|\mathscr{B})$ almost surely,*
(c) *$\|E(X|\mathscr{B})\|_1 \le \|X\|_1$,*
(d) *if X is $\mathscr{B}$-measurable, then $E(X|\mathscr{B}) = X$ almost surely (in particular, if c is a constant, then $E(c|\mathscr{B}) = c$ almost surely),*
(e) *if $\mathscr{B}_0 \subseteq \mathscr{B}$, then $E(X|\mathscr{B}_0) = E(E(X|\mathscr{B})|\mathscr{B}_0)$ almost surely,*
(f) *if $\mathscr{B}$ and X are independent (that is, if $\mathscr{B}$ and $\sigma(X)$ are independent), then $E(X|\mathscr{B})$ is almost surely equal to the constant $E(X)$, and*
(g) *if X is bounded and $\mathscr{B}$-measurable, then $E(XY|\mathscr{B}) = XE(Y|\mathscr{B})$ almost surely.*

Proof. Note that $aE(X|\mathscr{B}) + bE(Y|\mathscr{B})$ is a $\mathscr{B}$-measurable function that satisfies

[9] It is probably worth translating one of the parts of this proposition into more precise language. Part (a) says that if Z is a version of $E(aX + bY|\mathscr{B})$, if Z_1 is a version of $E(X|\mathscr{B})$, and if Z_2 is a version of $E(Y|\mathscr{B})$, then $Z = aZ_1 + bZ_2$ almost surely. Equivalently, part (a) can be viewed as saying that if Z_1 and Z_2 are versions of $E(X|\mathscr{B})$ and $E(Y|\mathscr{B})$, then $aZ_1 + bZ_2$ is a version of $E(aX + bY|\mathscr{B})$. Other assertions about conditional expectations can be made precise in similar ways.

$$\int_B (aE(X|\mathscr{B}) + bE(Y|\mathscr{B}))\,dP = \int_B (aX + bY)\,dP$$

for each B in $\mathscr{B}$ and hence is a conditional expectation of $aX + bY$ given $\mathscr{B}$. Part (a) then follows from the uniqueness of conditional expectations (part (b) of Proposition 10.4.2).

For part (b), note that

$$\int_B E(X|\mathscr{B})\,dP = \int_B X\,dP \le \int_B Y\,dP = \int_B E(Y|\mathscr{B})\,dP$$

holds for each B in $\mathscr{B}$. It now follows from Corollary 2.3.13 that $E(X|\mathscr{B}) \le E(Y|\mathscr{B})$ almost surely.

If we let A_+ and A_- be the sets $\{E(X|\mathscr{B}) \ge 0\}$ and $\{E(X|\mathscr{B}) < 0\}$, then part (c) follows from the calculation

$$\begin{aligned}\|E(X|\mathscr{B})\|_1 &= \int_{A_+} E(X|\mathscr{B})\,dP - \int_{A_-} E(X|\mathscr{B})\,dP \\ &= \int_{A_+} X\,dP - \int_{A_-} X\,dP \le \|X\|_1.\end{aligned}$$

Part (d) is immediate, and part (e) follows from the calculation

$$\int_B E(E(X|\mathscr{B})|\mathscr{B}_0)\,dP = \int_B E(X|\mathscr{B})\,dP = \int_B X\,dP$$

which holds for every B in $\mathscr{B}_0$ (recall that $\mathscr{B}_0 \subseteq \mathscr{B}$).

We turn to part (f). If $\mathscr{B}$ and X are independent, then for each B in $\mathscr{B}$ the random variables χ_B and X are independent, and so Proposition 10.1.10 implies that

$$\begin{aligned}\int_B X\,dP &= \int \chi_B X\,dP = \int \chi_B\,dP \int X\,dP \\ &= P(B)E(X) = \int_B E(X)\,dP;\end{aligned}$$

it follows that $E(X)$ is a version of $E(X|\mathscr{B})$.

Let us start our consideration of part (g) with the special case where $X = \chi_A$ for some A in $\mathscr{B}$. Then for each B in $\mathscr{B}$ we have

$$\begin{aligned}\int_B XY\,dP &= \int_{B\cap A} Y\,dP = \int_{B\cap A} E(Y|\mathscr{B})\,dP \\ &= \int_B \chi_A E(Y|\mathscr{B})\,dP = \int_B XE(Y|\mathscr{B})\,dP\end{aligned}$$

and so

$$\int_B XY\,dP = \int_B XE(Y|\mathscr{B})\,dP. \tag{3}$$

Equation (3) now extends to the case where X is simple function and then (by the dominated convergence theorem) to the case where X is an arbitrary bounded $\mathscr{B}$-measurable function. Furthermore $XE(Y|\mathscr{B})$ is $\mathscr{B}$-measurable. Thus $XE(Y|\mathscr{B})$ is a version of $E(XY|\mathscr{B})$ and the proof is complete. □

Proposition 10.4.4 (Monotone and Dominated Convergence Theorems for Conditional Expectations). *Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\mathscr{B}$ be a sub-σ-algebra of $\mathscr{A}$, and let $X_1, X_2, \ldots$ be random variables with finite expected values such that $\lim_n X_n$ exists almost surely. If*

(a) *$\{X_n\}$ is an increasing sequence such that $\lim_n E(X_n)$ is finite, or*
(b) *there exists a random variable Y with finite expected value such that each X_n satisfies $|X_n| \leq Y$ almost surely,*

then $\lim_n X_n$ has a finite expected value and $E(\lim_n X_n|\mathscr{B}) = \lim_n E(X_n|\mathscr{B})$ almost surely.

Proof. First suppose that condition (a) holds. Let us also temporarily assume that the random variables X_n are nonnegative. Since we are assuming that $\{X_n\}$ is an increasing sequence, it follows from part (b) of Proposition 10.4.3 that the sequence $\{E(X_n|\mathscr{B})\}$ is increasing almost surely and so has an almost sure limit, possibly with some of values of $\lim_n E(X_n|\mathscr{B})$ being infinite. The monotone convergence theorem implies that

$$\int \lim_n E(X_n|\mathscr{B})\, dP = \lim_n \int E(X_n|\mathscr{B})\, dP = \lim_n \int X_n\, dP < +\infty,$$

and so $\lim_n E(X_n|\mathscr{B})$ is finite almost everywhere. Applying the monotone convergence theorem twice more gives

$$\int_B \lim_n X_n\, dP = \lim_n \int_B X_n\, dP = \lim_n \int_B E(X_n|\mathscr{B})\, dP = \int_B \lim_n E(X_n|\mathscr{B})\, dP$$

for each B in $\mathscr{B}$; thus $\lim_n E(X_n|\mathscr{B})$ is a version of $E(\lim_n X_n|\mathscr{B})$ and the proof is complete in the case where condition (a) holds and the X_n's are nonnegative. We can complete the proof for the case where (a) holds by applying what we have just proved to the sequence $\{X_n - X_1\}$ and then using the linearity of conditional expectations.

Now suppose that condition (b) holds. Since we are assuming that $|X_n| \leq Y$ for each n, we have $|\lim_n X_n| \leq Y$ and so $\lim_n X_n$ has a finite expected value. For each n let $Y_n = \inf\{X_k : k \geq n\}$ and $Z_n = \sup\{X_k : k \geq n\}$. Then $\{Y_n\}$ is an increasing sequence that converges pointwise to $\liminf_n X_n$, and $\{Z_n\}$ is a decreasing sequence that converges pointwise to $\limsup_n X_n$; since $\lim X_n$ exists, both those sequences converge to it almost surely. If we apply the first half of the proposition to the sequence $\{Y_n\}$, we conclude that $\lim_n E(Y_n|\mathscr{B}) = E(\lim_n Y_n|\mathscr{B}) = E(\lim_n X_n|\mathscr{B})$ almost surely. A similar argument, applied to the sequence $\{Y - Z_n\}$, shows that $\lim_n E(Z_n|\mathscr{B}) = E(\lim_n X_n|\mathscr{B})$ almost surely. Finally, each variable $E(X_n|\mathscr{B})$ lies

between the corresponding variables $E(Y_n|\mathscr{B})$ and $E(Z_n|\mathscr{B})$, and it follows that $\lim_n E(X_n|\mathscr{B}) = E(\lim_n X_n|\mathscr{B})$ almost surely. With this the proof is complete. □

In the remainder of this chapter we will be looking at *stochastic processes*. A rather abstract definition might say that a stochastic process is an indexed family $\{X_t\}_{t\in T}$ of random variables, where T is an arbitrary nonempty set and all the random variables are defined on the same probability space. However, one usually deals with more concrete situations, in which the index set T is a set of integers or else a nice set of real numbers (such as an interval), and the members of T are interpreted as times. For each t in T the random variable X_t is thought of as representing a quantity that can be observed at time t.

A *discrete-time* stochastic process is one for which T is a set of integers, and a *continuous-time* process is one for which T is an interval of real numbers. We will see a few discrete-time processes in this section, and we will see some continuous-time processes later in the chapter.

Let $(\Omega, \mathscr{A}, P)$ be a probability space. A *filtration*[10] is a sequence $\{\mathscr{F}_n\}_{n=0}^{\infty}$ of sub-σ-algebras of $\mathscr{A}$ that is increasing, in the sense that $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$ holds for each n. A discrete-time stochastic process (i.e., a sequence of random variables) $\{X_n\}_{n=0}^{\infty}$ is *adapted* to the filtration $\{\mathscr{F}_n\}_{n=0}^{\infty}$ if X_n is $\mathscr{F}_n$-measurable for each n. Note that the sequence $\{X_n\}_{n=0}^{\infty}$ is adapted to the filtration $\{\mathscr{F}_n\}_{n=0}^{\infty}$ if and only if $\sigma(X_0, \ldots, X_n) \subseteq \mathscr{F}_n$ holds for each n.

The intuition here is that the events in the σ-algebra $\mathscr{F}_n$ are those that could be known by time n. In one common situation, $\{X_n\}$ is an arbitrary sequence of random variables and for each n we let $\mathscr{F}_n$ be $\sigma(X_0, \ldots, X_n)$. In this case $\mathscr{F}_n$ contains exactly the events that are determined by the random variables $X_0, \ldots, X_n$.

Let $\{\mathscr{F}_n\}$ be a filtration on the probability space $(\Omega, \mathscr{A}, P)$. A *stopping time* or an *optional time* is a function $\tau\colon \Omega \to \mathbb{N}_0 \cup \{+\infty\}$ such that $\{\tau \le n\} \in \mathscr{F}_n$ holds for each n in $\mathbb{N}_0$. It is easy to check that if τ is a stopping time, then τ is $\mathscr{A}$-measurable and that a function $\tau\colon \Omega \to \mathbb{N}_0 \cup \{+\infty\}$ is a stopping time if and only if $\{\tau = n\} \in \mathscr{F}_n$ holds for each n in $\mathbb{N}_0$.

One standard interpretation of a stopping time is the following: You are observing random variables X_0, X_1, ..., one after the other, and you may decide to stop observing at some random time τ. It is reasonable to decide whether or not to stop with the nth observation on the basis of the information that is available by time n, but it is not reasonable to use information about the future (e.g., the values of X_{n+1}, X_{n+2}, ...). In other words, $\{\tau = n\}$, the event that you stop just after observing X_n, should belong to $\mathscr{F}_n$.

---

[10] In this section we are dealing with discrete-time processes. On the other hand, a *filtration* $\{\mathscr{F}_t\}_{t\in T}$ in continuous time is defined by requiring that $\mathscr{F}_{t_1} \subseteq \mathscr{F}_{t_2}$ holds whenever t_1 and t_2 are elements of T such that $t_1 < t_2$. If $\{\mathscr{F}_t\}_{t\in T}$ is a filtration with $T = [0, +\infty)$, then a stopping time for it is a function $\tau\colon \Omega \to [0, +\infty]$ such that $\{\tau \le t\} \in \mathscr{F}_t$ holds for all t in T. Except for a few exercises involving Brownian motion, we will not be dealing with filtrations in continuous time.

Example 10.4.5. Suppose that you take a random walk on the integers in the following way. You begin at 0, and every minute you toss a fair coin and move to the right by a distance of 1 if the coin yields a head and to the left by a distance of 1 if it yields a tail. To formalize this, we let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed random variables such that

$$P(\{Y_i = -1\}) = P(\{Y_i = 1\}) = 1/2$$

holds for each i, and then we define $\{X_n\}_{n=0}^{\infty}$ by $X_0 = 0$ and $X_n = Y_1 + \cdots + Y_n$ if $n > 0$. Finally, we define the filtration $\{\mathscr{F}_n\}$ by letting $\mathscr{F}_n$ be $\sigma(X_0, \ldots, X_n)$ for each n.

Let us consider a rather simple stopping time for this process. The time you first reach 1 (if you ever reach it) is given by

$$\tau_{\{1\}}(\omega) = \inf\{n \in \mathbb{N}_0 : X_n(\omega) = 1\}. \tag{4}$$

Note that $\tau_{\{1\}}(\omega) = +\infty$ if the set on the right side of (4) is empty—in other words, if you never reach the point 1. Since

$$\{\tau_{\{1\}} \leq n\} = \bigcup_{i \leq n} \{X_i = 1\} \in \mathscr{F}_n,$$

the variable $\tau_{\{1\}}$ is in fact a stopping time. □

Example 10.4.6. Now suppose we have an arbitrary real-valued process $\{X_n\}_{n=0}^{\infty}$ that is adapted to some filtration $\{\mathscr{F}_n\}$ and we want to know the first time that X_n is in some Borel subset A of $\mathbb{R}$. The same reasoning as in Example 10.4.5 works if we replace (4) with

$$\tau_A(\omega) = \inf\{n \in \mathbb{N}_0 : X_n(\omega) \in A\}. \quad \square$$

Let us now turn to martingales. Suppose that $(\Omega, \mathscr{A}, P)$ is a probability space, that $\{\mathscr{F}_n\}_{n=0}^{\infty}$ is a filtration on $(\Omega, \mathscr{A}, P)$, and that $\{X_n\}_{n=0}^{\infty}$ is a discrete-time process on $(\Omega, \mathscr{A}, P)$. Then $(\{X_n\}_{n=0}^{\infty}, \{\mathscr{F}_n\}_{n=0}^{\infty})$, or simply $\{X_n\}_{n=0}^{\infty}$, is a *martingale* if

(a) $\{X_n\}_{n=0}^{\infty}$ is adapted to $\{\mathscr{F}_n\}_{n=0}^{\infty}$,
(b) each X_n has a finite expected value, and
(c) for each n we have $X_n = E(X_{n+1}|\mathscr{F}_n)$ almost surely.

Sometimes we will say that $\{X_n\}$ is a martingale *relative to* $\{\mathscr{F}_n\}$. If condition (c) is replaced with

$$\text{for each } n \text{ we have } X_n \leq E(X_{n+1}|\mathscr{F}_n) \text{ almost surely}$$

or with

$$\text{for each } n \text{ we have } X_n \geq E(X_{n+1}|\mathscr{F}_n) \text{ almost surely},$$

then $(\{X_n\}_{n=0}^{\infty}, \{\mathscr{F}_n\}_{n=0}^{\infty})$ or $\{X_n\}_{n=0}^{\infty}$ is a *submartingale* or a *supermartingale*. Note that we can verify condition (c) in the definition of a martingale by checking that $\int_B X_n\,dP = \int_B X_{n+1}\,dP$ holds for each n in $\mathbb{N}_0$ and each B in $\mathscr{F}_n$. Similar remarks apply to submartingales and supermartingales.

Examples 10.4.7.

(a) Let $(\Omega, \mathscr{A}, P)$ be a probability space, and let $\{Y_n\}_{n=1}^{\infty}$ be a sequence of independent (real-valued) random variables on Ω with finite expectations. Define $\{S_n\}_{n=0}^{\infty}$ by $S_0 = 0$ and $S_n = Y_1 + \cdots + Y_n$ if $n \geq 1$, and define a filtration $\{\mathscr{F}_n\}_{n=0}^{\infty}$ by $\mathscr{F}_n = \sigma(S_0, \ldots, S_n)$. If $E(Y_n) = 0$ for $n = 1, 2, \ldots$, then we can use parts (a), (d), and (f) of Proposition 10.4.3, together with the independence of the sequence $\{Y_n\}_{n=1}^{\infty}$, to show that

$$E(S_{n+1}|\mathscr{F}_n) = E(S_n + Y_{n+1}|\mathscr{F}_n) = S_n + E(Y_{n+1}|\mathscr{F}_n) = S_n$$

holds almost surely for each n, and hence that $\{S_n\}_{n=0}^{\infty}$ is a martingale. Similar calculations show that if $E(Y_n) \geq 0$ for $n = 1, 2, \ldots$ (or if $E(Y_n) \leq 0$ for $n = 1, 2, \ldots$), then $\{S_n\}_{n=0}^{\infty}$ is a submartingale (or a supermartingale).

(b) Suppose that you are gambling, making a sequence of wagers. Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of random variables with finite expected values and defined on some probability space $(\Omega, \mathscr{A}, P)$, and suppose that X_0 represents your capital at the start and that X_n represents your capital after n wagers. Define a filtration by letting $\mathscr{F}_n = \sigma(X_0, \ldots, X_n)$ hold for each n. Then $\{X_n\}_{n=0}^{\infty}$ is a martingale if the wagers are fair (that is, if at each stage the conditional expectation of your gain from the next wager, namely $E(X_{n+1}|\mathscr{F}_n) - X_n$, is 0); it is a submartingale if the wagers favor you and is a supermartingale if they favor your opponent.

(c) Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\{\mathscr{F}_n\}_{n=0}^{\infty}$ be a filtration on $(\Omega, \mathscr{A}, P)$, and let X be an integrable $\mathscr{A}$-measurable function on Ω. For each n define X_n by $X_n = E(X|\mathscr{F}_n)$. Let us check that $\{X_n\}_{n=0}^{\infty}$ is a martingale. Condition (c) in the definition of martingales is the only thing to check, and that condition follows from the calculation

$$E(X_{n+1}|\mathscr{F}_n) = E(E(X|\mathscr{F}_{n+1})|\mathscr{F}_n) = E(X|\mathscr{F}_n) = X_n$$

(see part (e) of Proposition 10.4.3).

(d) We define a martingale on the probability space $((0,1], \mathscr{B}((0,1]), \lambda)$ as follows. Let $\mathscr{F}_0$ be the σ-algebra that contains only the sets $\varnothing$ and $(0,1]$. For positive n let $\mathscr{P}_n$ be the partition of $(0,1]$ that consists of the intervals $(i/2^n, (i+1)/2^n]$, $i = 0, \ldots, 2^n - 1$; then let $\mathscr{F}_n = \sigma(\mathscr{P}_n)$. Now suppose that μ is a finite Borel measure on $(0,1]$, and for each n define $X_n\colon (0,1] \to \mathbb{R}$ by $X_n(x) = \mu(I)/\lambda(I)$, where I is the interval in $\mathscr{P}_n$ that contains x. Then each interval I in $\mathscr{P}_n$ satisfies

$$\int_I X_n\,d\lambda = \mu(I) = \int_I X_{n+1}\,d\lambda.$$

It follows that the same equation holds if I is replaced with an arbitrary set in $\mathscr{F}_n$; hence $X_n = E(X_{n+1}|\mathscr{F}_n)$ and $\{X_n\}$ is a martingale. There are a couple of things to note here. First, if we consider the behavior of the sequence $\{X_n(x)\}$ as n goes to infinity, we seem to be dealing with some sort of derivative. We'll look harder at this later in this section. Second, we are dealing with pure analysis in this example; no probability seems to be involved. □

The following is one of the major results of martingale theory.

Theorem 10.4.8 (Doob's Martingale Convergence Theorem). *Let $(\Omega, \mathscr{A}, P)$ be a probability space, and let $(\{X_n\}_{n=0}^{\infty}, \{\mathscr{F}_n\}_{n=0}^{\infty})$ be a submartingale on $(\Omega, \mathscr{A}, P)$ such that $\sup_n E(X_n^+) < +\infty$. Then the limit $\lim_n X_n$ exists almost surely, and $E(|\lim_n X_n|) < +\infty$.*

We need a few preliminary results before we prove the martingale convergence theorem.

Lemma 10.4.9. *Suppose that $\{\mathscr{F}_n\}$ is a filtration on the probability space $(\Omega, \mathscr{A}, P)$ and that $\{X_n\}$ and $\{Y_n\}$ are submartingales on Ω relative to $\{\mathscr{F}_n\}$. Then $\{X_n \vee Y_n\}$ is a submartingale relative to $\{\mathscr{F}_n\}$.*

Proof. It is clear that each $X_n \vee Y_n$ has a finite expectation and is $\mathscr{F}_n$-measurable. Define sets C_n, $n = 0, 1, \ldots$, by $C_n = \{X_n > Y_n\}$. Then each C_n belongs to the corresponding $\mathscr{F}_n$, and for each B in $\mathscr{F}_n$ we have

$$\begin{aligned}\int_B (X_n \vee Y_n)\, dP &= \int_{B\cap C_n} X_n\, dP + \int_{B\cap C_n^c} Y_n\, dP \\ &\leq \int_{B\cap C_n} X_{n+1}\, dP + \int_{B\cap C_n^c} Y_{n+1}\, dP \leq \int_B (X_{n+1} \vee Y_{n+1})\, dP.\end{aligned}$$

Thus $\{X_n \vee Y_n\}$ is a submartingale relative to $\{\mathscr{F}_n\}$. □

Let us for a moment view a martingale (or sub- or supermartingale) $\{X_n\}$ in terms of gambling, with X_n representing our capital after the nth of a sequence of games. It is sometimes useful to modify $\{X_n\}$ by allowing ourselves to skip certain of the games. More precisely, let $\{\varepsilon_n\}$ be a sequence of $\{0,1\}$-valued random variables, with ε_n having value 1 if we participate in the nth game and having value 0 otherwise. Since $X_n - X_{n-1}$ would be our gain or loss from the nth game of the original sequence, $\varepsilon_n(X_n - X_{n-1})$ will be our gain or loss in the modified sequence. Thus we can describe our fortunes in the modified situation with a sequence $\{Y_n\}$, where $Y_0 = X_0$ and $Y_n = Y_{n-1} + \varepsilon_n(X_n - X_{n-1})$, or, equivalently, $Y_n = X_0 + \sum_{i=1}^n \varepsilon_i(X_i - X_{i-1})$. For this formalization to be reasonable, we must make our decisions about which games to play and which to skip using only information that is available at the time of the decision. Hence it is natural to assume that ε_n is $\mathscr{F}_{n-1}$-measurable.

We have the following proposition, which says that if we transform a submartingale $\{X_n\}$ as in the preceding paragraph, then the resulting sequence $\{Y_n\}$ is also a submartingale, with expected values no larger than those for the original sequence.

Proposition 10.4.10. *Suppose that $(\{X_n\},\{\mathscr{F}_n\})$ is a submartingale on the probability space $(\Omega,\mathscr{A},P)$ and that $\{\varepsilon_n\}_{n=1}^{\infty}$ is a sequence of $\{0,1\}$-valued random variables on Ω such that ε_n is $\mathscr{F}_{n-1}$-measurable for each n. Then the sequence $\{Y_n\}_{n=0}^{\infty}$ defined by $Y_0 = X_0$ and $Y_n = Y_{n-1} + \varepsilon_n(X_n - X_{n-1})$ for $n = 1, 2, \ldots$ is a submartingale, and $E(Y_n) \le E(X_n)$ holds for each n.*

Proof. It is clear that each Y_n is $\mathscr{F}_n$-measurable and has a finite expected value. Since $\{X_n\}$ is a submartingale,

$$E(X_n - X_{n-1}|\mathscr{F}_{n-1}) = E(X_n|\mathscr{F}_{n-1}) - X_{n-1} \ge 0$$

holds almost surely for $n = 1, 2, \ldots$, and so (see Proposition 10.4.3)

$$\begin{aligned} E(Y_n|\mathscr{F}_{n-1}) &= E(Y_{n-1}|\mathscr{F}_{n-1}) + E(\varepsilon_n(X_n - X_{n-1})|\mathscr{F}_{n-1}) \\ &= Y_{n-1} + \varepsilon_n E(X_n - X_{n-1}|\mathscr{F}_{n-1}) \\ &\ge Y_{n-1} \end{aligned}$$

almost surely; thus $\{Y_n\}$ is a submartingale. We prove that $E(Y_n) \le E(X_n)$ by induction. This inequality certainly holds when $n = 0$. For the induction step, note that, since $E(X_n - X_{n-1}|\mathscr{F}_{n-1}) \ge 0$, we have

$$\begin{aligned} E(Y_n) &= E(Y_{n-1}) + E(\varepsilon_n(X_n - X_{n-1})) \\ &= E(Y_{n-1}) + E(\varepsilon_n E(X_n - X_{n-1}|\mathscr{F}_{n-1})) \\ &\le E(X_{n-1}) + E(X_n - X_{n-1}) = E(X_n). \end{aligned}$$

□

In order to prove the martingale convergence theorem, we will look a bit at how a sequence $\{x_n\}$ of real numbers might fail to converge. One way for this to happen is for $\liminf_n x_n$ to be less than $\limsup_n x_n$. In that case, there are real numbers a and b such that

$$\liminf_n x_n < a < b < \limsup_n x_n,$$

from which it follows that there is a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_1} < a$, $x_{n_2} > b$, $x_{n_3} < a, \ldots$. This suggests the following definition. A sequence $\{x_n\}$ is said to have an *upcrossing* of the interval $[a,b]$ as n increases from p to q if $x_p \le a$, $x_n < b$ for n satisfying $p < n < q$, and $x_q \ge b$.

Now suppose that $(\Omega,\mathscr{A},P)$ is a probability space, that $\{\mathscr{F}_n\}$ is a filtration on $(\Omega,\mathscr{A},P)$, and that $\{X_n\}_{n=0}^{\infty}$ is a sequence of random variables adapted to $\{\mathscr{F}_n\}$. Let a and b be real numbers such that $a < b$. Our immediate goal is to count the upcrossings of the interval $[a,b]$ made by these random variables, and for this we use sequences $\{\sigma_n\}$ and $\{\tau_n\}$ of stopping times defined as follows. We define σ_1 by

$$\sigma_1(\omega) = \inf\{i \in \mathbb{N}_0 \colon X_i(\omega) \le a\},$$

and then we continue inductively, defining σ_n, $n \ge 2$, and τ_n, $n \ge 1$, by

$$\tau_n(\omega) = \inf\{i \in \mathbb{N}_0 \colon i > \sigma_n(\omega) \text{ and } X_i(\omega) \ge b\}$$

and

$$\sigma_n(\omega) = \inf\{i \in \mathbb{N}_0 \colon i > \tau_{n-1}(\omega) \text{ and } X_i(\omega) \le a\}$$

(recall that the infimum of the empty set is $+\infty$). We can check inductively that σ_n and τ_n are indeed stopping times by noting that

$$\{\sigma_1 \le k\} = \cup_{i=0}^{k} \{X_i \le a\} \in \mathscr{F}_k,$$

$$\{\sigma_n \le k\} = \cup_{i=1}^{k} \{\tau_{n-1} < i \text{ and } X_i \le a\} \in \mathscr{F}_k \quad \text{if } n \ge 2, \text{ and}$$

$$\{\tau_n \le k\} = \cup_{i=1}^{k} \{\sigma_n < i \text{ and } X_i \ge b\} \in \mathscr{F}_k.$$

The finite sequence $\{X_i(\omega)\}_{i=0}^{n}$ contains k or more upcrossings[11] of $[a,b]$ if and only if $\tau_k(\omega) \le n$. Thus, if we define functions $U_n^{[a,b]} \colon \Omega \to \mathbb{R}$ by letting $U_n^{[a,b]}(\omega)$ be the number of upcrossings of $[a,b]$ in the sequence $\{X_i(\omega)\}_{i=0}^{n}$, then $\{U_n^{[a,b]} \ge k\} = \{\tau_k \le n\}$; since each τ_k is a stopping time, it follows that $U_n^{[a,b]}$ is $\mathscr{F}_n$-measurable.

Proposition 10.4.11 (The upcrossing inequality). *Let $(\Omega, \mathscr{A}, P)$ be a probability space and let $(\{X_n\}, \{\mathscr{F}_n\})$ be a submartingale on $(\Omega, \mathscr{A}, P)$. If a and b are real numbers such that $a < b$, then for each n the number $U_n^{[a,b]}$ of upcrossings of $[a,b]$ by $\{X_i\}_{i=0}^{n}$ satisfies*

$$E(U_n^{[a,b]}) \le \frac{E((X_n - a)^+)}{b - a}.$$

Proof. Let us suppose that a and b are fixed. We can assume that each X_n satisfies $a \le X_n$, since replacing $\{X_n\}$ with $\{\max(X_n, a)\}$ gives a new sequence that is a submartingale (see Lemma 10.4.9), has the same number of upcrossings of $[a,b]$ as the original sequence, and is such that $E((X_n - a)^+)$ is the same for the old and new sequences. Let $\{\sigma_n\}$ and $\{\tau_n\}$ be the sequences of stopping times defined before the statement of the proposition, and define functions[12] $\varepsilon_n \colon \Omega \to \mathbb{R}$, $n = 1, 2, \ldots$, by

$$\varepsilon_n(\omega) = \begin{cases} 1 & \text{if there is an } i \text{ such that } \sigma_i(\omega) < n \le \tau_i(\omega), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Then

[11] Here we are, of course, counting non-overlapping upcrossings, where we call a sequence of upcrossings of $[a,b]$ non-overlapping if the sets of times (i.e., of subscripts) during which they occur are non-overlapping.

[12] The intuitive meaning of ε_n is that it tells whether X_n is part of an upcrossing.

$$\{\varepsilon_n = 1\} = \cup_i (\{\sigma_i \le n-1\} \cap \{\tau_i \le n-1\}^c) \in \mathscr{F}_{n-1},$$

and so ε_n is $\mathscr{F}_{n-1}$-measurable. Let $\{Y_n\}$ be the submartingale (see Proposition 10.4.10) defined by $Y_n = X_0 + \sum_{i=1}^n \varepsilon_i (X_i - X_{i-1})$. We will use $\{Y_n\}$ to bound the number of upcrossings of $[a,b]$ by $\{X_i\}_{i=0}^n$.

For an arbitrary element ω of Ω let us analyze the set of those k that satisfy $k \le n$ and $\varepsilon_k(\omega) = 1$. Such values of k can arise in two ways. First, for each i such that $\tau_i(\omega) \le n$ we have the set of k that satisfy $\sigma_i(\omega) < k \le \tau_i(\omega)$. Those values correspond to the steps in the upcrossing of $[a,b]$ that begins at $\sigma_i(\omega)$ and ends at $\tau_i(\omega)$, and so we have

$$b - a \le \sum_{k=\sigma_i(\omega)+1}^{\tau_i(\omega)} (X_k(\omega) - X_{k-1}(\omega)). \tag{5}$$

The other way that such k can arise is for there to be an i such that $\sigma_i(\omega) < k \le n < \tau_i(\omega)$. These k correspond to a potential upcrossing that has started but has not finished by time n, and in this case we have

$$\sum_{k=\sigma_i(\omega)+1}^{n} (X_k(\omega) - X_{k-1}(\omega)) = X_n(\omega) - a \ge 0. \tag{6}$$

We are now ready to relate the number of upcrossings to the submartingale $\{Y_n\}$. In view of (5) and (6), we have

$$X_0 + (b-a)U_n^{[a,b]} \le X_0 + \sum_{k=1}^n \varepsilon_k (X_k - X_{k-1}) = Y_n;$$

since $a \le X_0$ and $E(Y_n) \le E(X_n)$ (see Proposition 10.4.10), it follows that

$$a + (b-a)E(U_n^{[a,b]}) \le E(Y_n) \le E(X_n)$$

and hence that

$$(b-a)E(U_n^{[a,b]}) \le E(X_n - a) \le E((X_n - a)^+).$$

With this the proof of the upcrossing lemma is complete. □

We are now in a position to prove the martingale convergence theorem.

Proof of the Martingale Convergence Theorem. As in the statement of the theorem, let $\{X_n\}_{n=0}^\infty$ be a submartingale such that $\sup_n E(X_n^+) < +\infty$. We begin by showing that $\liminf_n X_n = \limsup_n X_n$ almost surely, which we do by counting upcrossings.

For each pair a, b of real numbers such that $a < b$ we define $U^{[a,b]} : \Omega \to \mathbb{R}$ by letting $U^{[a,b]}(\omega)$ be the total number of upcrossings of $[a,b]$ in the sequence $\{X_n(\omega)\}_{n=0}^\infty$. (This differs from $U_n^{[a,b]}$, which only counts the upcrossings in the first

$n+1$ terms of $\{X_i(\omega)\}_{i=0}^{\infty}$). Note that the sequence $\{U_n^{[a,b]}\}_{n=1}^{\infty}$ is increasing and has $U^{[a,b]}$ as its limit, and also that $(X_n-a)^+ \le X_n^+ + |a|$. The monotone convergence theorem and the upcrossing inequality, together with assumption that $\sup_n E(X_n^+) < +\infty$, imply that

$$E(U^{[a,b]}) = \lim_n E(U_n^{[a,b]}) \le \sup_n \frac{E((X_n-a)^+)}{b-a} \le \frac{\sup_n EX_n^+ + |a|}{b-a} < +\infty.$$

It follows that $U^{[a,b]}$, the number of upcrossings of $[a,b]$, is almost surely finite. Since

$$\{\liminf_n X_n < \limsup_n X_n\} = \cup_{a,b}\{U^{[a,b]} = +\infty\},$$

where a and b range over all rational numbers such that $a<b$, we have $\liminf_n X_n = \limsup_n X_n$ almost surely. Thus $\lim_n X_n$ exists almost surely, as an element of $[-\infty,+\infty]$. We still need to show that $E(|\lim_n X_n|) < +\infty$ and hence that $\lim_n X_n$ is finite almost surely.

Note that $|X_n| = 2X_n^+ - X_n$, and so if we use Fatou's lemma (Theorem 2.4.4), plus the fact that $\{X_n\}$, as a submartingale, satisfies $E(X_0) \le E(X_n)$, we find

$$\begin{aligned}\int |\lim_n X_n|\,dP &= \int \liminf_n |X_n|\,dP \\ &\le \liminf_n \int |X_n|\,dP \le 2\sup_n \int X_n^+ - \int X_0\,dP < +\infty.\end{aligned}$$

With this the proof of the martingale convergence theorem is complete. □

Let us return to a couple of the examples discussed above. We first look at Example 10.4.7(c), which we can extend as follows:

Proposition 10.4.12. *Let $(\Omega,\mathscr{A},P)$ be a probability space, let X be an integrable random variable on Ω, let $\{\mathscr{F}_n\}$ be a filtration on $(\Omega,\mathscr{A},P)$, and let $\mathscr{F}_\infty = \sigma(\cup_n \mathscr{F}_n)$. Then the martingale $\{X_n\}$ defined by $X_n = E(X|\mathscr{F}_n)$ converges almost surely and in mean (i.e., in the norm $\|\cdot\|_1$) to $E(X|\mathscr{F}_\infty)$.*

Proof. Since

$$E(X_n^+) = \int_{\{X_n\ge 0\}} X_n = \int_{\{X_n\ge 0\}} X \le \|X\|_1,$$

the martingale convergence theorem (Theorem 10.4.8) implies that the sequence $\{X_n\}$ converges almost surely, say to $X_{\lim}$.

Let $X_\infty = E(X|\mathscr{F}_\infty)$. Part (e) of Proposition 10.4.3 implies that $\{X_n\}$ is also given by $X_n = E(X_\infty|\mathscr{F}_n)$. Let us show that $\lim_n \|X_n - X_\infty\|_1 = 0$. Suppose that ε is a positive real number. It follows from Proposition 3.4.2 and Lemma 3.4.6 that there is a simple function X_ε of the form $\sum_i a_i \chi_{A_i}$, where each A_i belongs to $\cup_n \mathscr{F}_n$, such that $\|X_\varepsilon - X_\infty\|_1 < \varepsilon$. Since each A_i is in $\mathscr{F}_n$ for some n, there is a positive integer N such that X_ε is $\mathscr{F}_N$-measurable. It follows that if $n \ge N$, then $E(X_\varepsilon|\mathscr{F}_n) = X_\varepsilon$, and so (see also part (c) of Proposition 10.4.3)

$$\begin{aligned}\|X_n - X_\infty\|_1 &= \|E(X_\infty|\mathscr{F}_n) - X_\infty\|_1 \\ &\le \|E(X_\infty|\mathscr{F}_n) - E(X_\varepsilon|\mathscr{F}_n)\|_1 + \|E(X_\varepsilon|\mathscr{F}_n) - X_\varepsilon\|_1 + \|X_\varepsilon - X_\infty\|_1 \\ &\le \|X_\varepsilon - X_\infty\|_1 + 0 + \|X_\varepsilon - X_\infty\|_1 \le 2\varepsilon.\end{aligned}$$

Since ε was arbitrary, we have $\lim_n \|X_n - X_\infty\|_1 = 0$.

We still need to show that $\{X_n\}$ converges to X_∞ almost surely. Since we have $\lim_n \|X_n - X_\infty\|_1 = 0$, there is a subsequence of $\{X_n\}$ that converges to X_∞ almost surely (see the discussion that follows the proof of Proposition 3.1.5). Since we already know that the sequence $\{X_n\}$ converges almost surely to $X_{\lim}$, we can conclude that $X_\infty = X_{\lim}$ and hence that $\{X_n\}$ converges to X_∞ both almost surely and with respect to $\|\cdot\|_1$. □

See Exercise 11 for another proof of Proposition 10.4.12.

Example 10.4.13. Let us now look at Example 10.4.7(d), which hinted at some relationships between martingales and derivatives. Let μ be the measure from that example, and define F by $F(x) = \mu((0,x])$. The martingale convergence theorem says that the limit

$$\lim_n \frac{F(b_n) - F(a_n)}{b_n - a_n}$$

exists for almost every x in $(0,1]$, where for each n we let $(a_n, b_n]$ be the interval in $\mathscr{P}_n$ that contains x. In case μ is absolutely continuous with respect to Lebesgue measure, Proposition 10.4.12 identifies this limit as the Radon–Nikodym derivative of μ with respect to Lebesgue measure. See Exercise 12 for the case of singular measures.

Note that the argument in the preceding paragraph is not a derivation of the almost everywhere differentiability of monotone functions from the martingale convergence theorem—there are uncountably many possible choices for the sequence $\{\mathscr{P}_n\}$ of partitions of $(0,1]$, and different sequences of partitions could give rise to different sets of values where the limit does not exist. Nevertheless, as noted by Doob [38, p. 347], these ideas can be made to work; see Chatterji [27] for the details. □

Exercises

1. Let $(\Omega, \mathscr{A}, P)$ be a probability space, let X and Y be random variables on $(\Omega, \mathscr{A}, P)$ such that the joint distribution of (X,Y) on $\mathbb{R}^2$ is absolutely continuous with respect to Lebesgue measure, and let $p\colon \mathbb{R}^2 \to \mathbb{R}$ be the density function for that joint distribution. Suppose that $F\colon \mathbb{R}^2 \to \mathbb{R}$ is a Borel measurable function such that $F \circ (X,Y)$ has a finite expected value. Define a

function $f\colon \mathbb{R} \to \mathbb{R}$ by letting

$$f(x) = \frac{\int F(x,y)p(x,y)dy}{\int p(x,y)dy}$$

for those x for which the expression above is defined and finite and by letting $f(x) = 0$ for other x. Show that $f \circ X$ is a version of the conditional expectation $E(F \circ (X,Y)|\sigma(X))$.

2. Suppose that $(\Omega, \mathscr{A}, P)$ is a probability space and that $\{\mathscr{F}_n\}$ is a filtration on $(\Omega, \mathscr{A}, P)$.
   (a) Show that if τ_1 and τ_2 are stopping times and n is a positive integer, then $\tau_1 + n$, $\tau_1 + \tau_2$, $\tau_1 \vee \tau_2$, and $\tau_1 \wedge \tau_2$ are stopping times.
   (b) Show that if $\{\tau_n\}$ is a sequence of stopping times, then $\inf_n \tau_n$, $\sup_n \tau_n$, $\liminf_n \tau_n$, and $\limsup_n \tau_n$ are stopping times.
3. Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\{\mathscr{F}_n\}_0^\infty$ be a filtration on $(\Omega, \mathscr{A}, P)$, and let τ be a stopping time. Define $\mathscr{F}_\tau$ to be the set of all sets A in $\sigma(\cup \mathscr{F}_n)$ such that $A \cap \{\tau \leq n\} \in \mathscr{F}_n$ holds for each nonnegative integer n.
   (a) Show that $\mathscr{F}_\tau$ is a sub-σ-algebra of $\mathscr{A}$.
   (b) Show that a set A belongs to $\mathscr{F}_\tau$ if and only if it satisfies $A \cap \{\tau = n\} \in \mathscr{F}_n$ for each nonnegative integer n, along with $A \cap \{\tau = +\infty\} \in \sigma(\cup \mathscr{F}_n)$.
4. Suppose that $\{X_n\}_1^\infty$ is a sequence of independent identically distributed random variables on $(\Omega, \mathscr{A}, P)$. Define a filtration $\{\mathscr{F}_n\}_0^\infty$ by $\mathscr{F}_0 = \{\varnothing, \Omega\}$ and $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$ for $n = 1, 2, \ldots$. Suppose that τ is a stopping time such that $P(\tau < +\infty) = 1$. Define a sequence $\{Y_n\}$ of random variables by

$$Y_n(\omega) = \begin{cases} X_{\tau+n}(\omega) & \text{if } \tau(\omega) < +\infty, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Show that the random variables $\{Y_n\}$ are independent and identically distributed, with the same distributions as the X_n's. (Hint: Consider the probabilities of events of the form $\{\tau = m\} \cap \{Y_1 \in A_1\} \cap \{Y_2 \in A_2\} \cap \cdots \cap \{Y_n \in A_n\}$.)
   (b) Show that the σ-algebra $\mathscr{F}_\tau$ and the process $\{Y_n\}$ are independent. That is, show that the σ-algebras $\mathscr{F}_\tau$ and $\sigma(Y_n, n = 1, 2, \ldots)$ are independent.
5. (Jensen's inequality for conditional expectations) Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function, let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\mathscr{B}$ be a sub-σ-algebra of $\mathscr{A}$, and let X be a random variable on $(\Omega, \mathscr{A}, P)$ such that both X and $\varphi \circ X$ have finite expected values. Show that $\varphi \circ E(X|\mathscr{B}) \leq E(\varphi \circ X|\mathscr{B})$ holds almost surely. (Hint: Use ideas from Exercise 3.3.8 to show that there is a family $\mathscr{F}$ of functions, each of the form $x \mapsto ax + b$, such that

$$\varphi(x) = \sup\{f(x) : f \in \mathscr{F}\}$$

holds for each x in $\mathbb{R}$ and such that $f \circ E(X|\mathscr{B}) \leq E(\varphi \circ X|\mathscr{B})$ holds almost surely for each f in $\mathscr{F}$. To conclude that $\varphi \circ E(X|\mathscr{B}) \leq E(\varphi \circ X|\mathscr{B})$ holds almost surely, choose a countable subset $\mathscr{F}_0$ of $\mathscr{F}$ such that φ is the pointwise supremum of the functions in $\mathscr{F}_0$. (Why do we need $\mathscr{F}_0$ to be countable?) The existence of such a subset can be derived from item D.11 in the appendices.)

6. Show that if $\{X_n\}$ is a submartingale relative to $\{\mathscr{F}_n\}$, then it is a submartingale relative to $\{\sigma(X_0, X_1, \ldots, X_n)\}$.
7. Show that if $(\{X_n\}, \{\mathscr{F}_n\})$ is a submartingale and if τ is a stopping time, then $(\{X_{\tau \wedge n}\}, \{\mathscr{F}_n\})$ is a submartingale.
8. (This exercise has nothing to do with martingales or conditional expectations. It appears here as preparation for Exercise 10.) Suppose that $\{a_n\}$ is a sequence of real numbers such that the sequence $\{e^{ita_n}\}$ is convergent for all t in some Lebesgue measurable set of positive measure.
   (a) Show that $\{e^{ita_n}\}$ is convergent for all real t. (Hint: Use Proposition 1.4.10.)
   (b) Show that $\{a_n\}$ is convergent. (Hint: Choose an interval $[b,c]$ such that $\int_b^c \lim e^{ita_n}\, dt \neq 0$. Then consider the sequence $\{\int_b^c e^{ita_n}\, dt\}$.)
9. Suppose that $\{X_n\}$ is a sequence of independent random variables on some probability space. For each n define $\mathscr{F}_n$ and S_n by $\mathscr{F}_n = \sigma(X_1, X_2, \ldots, X_n)$ and $S_n = X_1 + X_2 + \cdots + X_n$. Suppose that t is a real number such that $\lim_n E(e^{itS_n})$ exists and is not equal to 0. Check that for such t we have $E(e^{itS_n}) \neq 0$ for all n. Let $Y_n = e^{itS_n}/E(e^{itS_n})$ for each n.
   (a) Verify that $(\{Y_n\}, \{\mathscr{F}_n\})$ is a martingale.
   (b) Conclude that the sequence $\{e^{itS_n}\}$ is almost surely convergent.
10. Let $\{X_n\}$ be a sequence of independent random variables, let $\sum_n X_n$ be the corresponding infinite series, let μ_n, $n = 1, 2, \ldots$ be the distributions of the partial sums of the series, and let ϕ_{μ_n}, $n = 1, 2, \ldots$ be the corresponding characteristic functions. Consider the following conditions:

    (i) The series $\sum_n X_n$ converges almost everywhere.
    (ii) The series $\sum_n X_n$ converges in probability.
    (iii) The series $\sum_n X_n$ converges in distribution (that is, the sequence $\{\mu_n\}$ converges in distribution to some probability measure).
    (iv) The sequence of characteristic functions $\{\phi_{\mu_n}\}$ has a nonzero pointwise limit on a set of positive measure. That is, $\lim_n \phi_{\mu_n}(t)$ exists and is nonzero for all t in some set of positive measure.

    We have seen that condition (i) implies condition (ii), condition (ii) implies condition (iii), and condition (iii) implies condition (iv) (see Proposition 3.1.2, Exercise 10.3.7, and Proposition 10.3.15). Now prove that condition (iv) implies condition (i). (Hint: Use Exercises 8 and 9.)
11. Let $(\{X_n\}, \{\mathscr{F}_n\})$ be a martingale on $(\Omega, \mathscr{A}, P)$ such that the sequence $\{X_n\}$ is uniformly integrable. (See Exercises 4.2.12–4.2.16.)
    (a) Show that $\{X_n\}$ converges almost surely and in mean to some random variable X.
    (b) Show that for each n the equality $X_n = E(X|\mathscr{F}_n)$ holds almost surely.

12. Suppose that μ is a finite measure on $((0,1],\mathscr{B}((0,1]))$ and that $\{X_n\}$ is the martingale defined in Example 10.4.7(d). Show that if μ is singular with respect to Lebesgue measure, then $\lim_n X_n = 0$ holds λ-almost everywhere on $(0,1]$.
13. Let $(\Omega,\mathscr{A},P)$ be a probability space. In this exercise we consider sequences $\{X_n\}$ and $\{\mathscr{F}_n\}$ that are indexed by the *negative* integers. The pair $(\{X_n\},\{\mathscr{F}_n\})$ is called a *reverse martingale* if

    (i) each $\mathscr{F}_n$ is a sub-σ-algebra of $\mathscr{A}$,
    (ii) $\mathscr{F}_m \subseteq \mathscr{F}_n$ holds whenever $m \le n$,
    (iii) each X_n is measurable with respect to the corresponding $\mathscr{F}_n$ and has a finite expected value, and
    (iv) $X_n = E(X_{n+1}|\mathscr{F}_n)$ holds for $n = -2, -3, \ldots$.

    Prove the convergence theorem for reverse martingales: if $(\{X_n\},\{\mathscr{F}_n\})$ is a reverse martingale, then there is a function $X_{-\infty}$ such that $X_{-\infty} = \lim_{n\to-\infty} X_n$ holds almost surely and in mean. Furthermore, $X_{-\infty} = E(X_{-1}|\cap_n \mathscr{F}_n)$. (Hint: Use the upcrossing inequality, and verify and use the fact that the sequence $\{X_n\}$ is uniformly integrable. See Exercises 4.2.12–4.2.16
14. In this exercise we derive the strong law of large numbers from the convergence theorem for reverse martingales (see Exercise 13). Suppose that $(\Omega,\mathscr{A},P)$ is a probability space and that $\{X_i\}$ is a sequence of independent identically distributed random variables on $(\Omega,\mathscr{A},P)$ that have finite expected values. For each positive integer n let $S_n = X_1 + X_2 + \cdots + X_n$ and define the σ-algebra $\mathscr{F}_{-n}$ to be $\sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$.
    (a) Let $\mathscr{F} = \sigma(S_n)$. Show that $E(X_1|\mathscr{F}) = E(X_2|\mathscr{F}) = \cdots = E(X_n|\mathscr{F})$ and conclude that $E(X_1|\mathscr{F}) = S_n/n$. (Hint: Using the map

    $$\omega \mapsto (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$$

    to convert this to a calculation on $\mathbb{R}^n$ might be useful.)
    (b) Show that $(\{S_n/n\},\{\mathscr{F}_n\})$ is a reverse martingale.
    (c) Use the convergence theorem for reverse martingales, together with Kolmogorov's zero–one law (see Exercise 10.2.2), to conclude that $\lim_n S_n/n = E(X_1)$ holds almost surely.

10.5 Brownian Motion

In this section we look at a continuous-time stochastic process that models Brownian motion, the random movement of a very small particle suspended in a fluid. Einstein seems to have been one of the first to study Brownian motion mathematically, and Norbert Wiener was the first to build a probability measure with which to describe Brownian motion. In fact, the basic probability measure defining a Brownian motion process is generally called a *Wiener measure*.

As we noted in Sect. 10.4, a continuous-time process is a stochastic process $\{X_t\}_{t\in T}$ for which the index set T is a reasonable subset of $\mathbb{R}$—typically an interval such as $[0,1]$ or $[0,+\infty)$. We will first construct a Brownian motion in which the index set is $[0,1]$ and then we'll note how to build one with index set $[0,+\infty)$.

Since one usually thinks of particles moving in three-dimensional space, it seems natural to construct a process $\{X_t\}_{t\in T}$ for which the variables X_t have values in $\mathbb{R}^3$. However, the trick of taking three independent one-dimensional process and using them to build a three-dimensional process works. More precisely, suppose that $\{X_t\}_{t\in T}$ is a one-dimensional Brownian motion on a probability space $(\Omega,\mathscr{A},P)$. Then it turns out that the three-dimensional process $\{X'_t\}_{t\in T}$ that is defined on the product of three copies of $(\Omega,\mathscr{A},P)$ by $X'_t((\omega_1,\omega_2,\omega_3)) = (X_t(\omega_1),X_t(\omega_2),X_t(\omega_3))$ has suitable properties. In any case, we will devote our attention to one-dimensional Brownian motion. We begin with a precise definition.

Suppose that $(\Omega,\mathscr{A},P)$ is a probability space and that T is either $[0,1]$ or $[0,+\infty)$. A stochastic process $\{X_t\}_{t\in T}$ with values in $\mathbb{R}$ is a *Brownian motion*[13] if

(a) $X_0(\omega) = 0$ for all ω in Ω,
(b) for each choice of $t_0, t_1, \ldots, t_n$ in T such that $t_0 < t_1 < \cdots < t_n$ the increments $X_{t_i} - X_{t_{i-1}}$, $i = 1, \ldots, n$, are independent, with $X_{t_i} - X_{t_{i-1}}$ having distribution $N(0, t_i - t_{i-1})$, that is, a normal distribution with mean 0 and variance $t_i - t_{i-1}$, and
(c) for each ω in Ω the function $X_\bullet(\omega)\colon T \to \mathbb{R}$ defined by $t \mapsto X_t(\omega)$ is continuous.

Given a process $\{X_t\}_{t\in T}$, the functions $t \mapsto X_t(\omega)$ are called the *paths* of the process. Thus condition (c) says that we are requiring the paths of a Brownian motion process to be continuous.

Theorem 10.5.1. *Let $T = [0,1]$. Then a one-dimensional Brownian motion with parameter set T exists. That is, there exist a probability space $(\Omega,\mathscr{A},P)$ and random variables X_t, $t \in T$, on Ω such that the stochastic process $\{X_t\}_{t\in T}$ is a Brownian motion.*

Proof. Let $(\Omega,\mathscr{A},P)$ be a probability space on which there exists a sequence $\{Z_n\}_{n=0}^{\infty}$ of independent random variables, each of which has a normal distribution with mean 0 and variance 1. (Recall that according to Corollary 10.1.16, such a sequence can be constructed on the probability space $([0,1],\mathscr{B}([0,1]),\lambda)$.) We will use such a sequence $\{Z_n\}$ to build a sequence of piecewise linear approximations to a Brownian motion process. More precisely, we will construct processes $\{X_t^n\}_{t\in T}$, $n = 0, 1, \ldots$, such that

(a′) for each n the paths of $\{X_t^n\}_{t\in T}$ satisfy $X_0^n(\omega) = 0$ for all ω and are piecewise linear, with the paths being linear on the intervals of the form $[(i-1)/2^n, i/2^n]$,

[13] Some authors only require conditions (a) and (c) in the definition of a Brownian motion to hold for all ω outside some P-null subset of Ω.

(b′) for each n the process $\{X_t^n\}_{t\in T}$, when restricted to the points $t_{i/2^n}$, $i = 0, \ldots, 2^n$, looks like a Brownian motion (that is, it has independent increments whose distributions are normal and have the required means and variances),
(c′) for almost every ω the sequence of functions $\{t \mapsto X_t^n(\omega)\}_{n=1}^{\infty}$ converges uniformly on $[0,1]$ as n approaches infinity, and
(d′) these processes satisfy $X_t^n(\omega) = X_t^{n+1}(\omega) = X_t^{n+2}(\omega) = \ldots$ for each n and ω and each t of the form $i/2^n$.

Now assume that we have constructed such a sequence of processes $\{X_t^n\}_{t\in T}$, and let A be an event of probability 1 such that if $\omega \in A$, then the sequence $\{t \mapsto X_t^n(\omega)\}_n$ converges uniformly on T. Define a process $\{X_t\}_{t\in T}$ by

$$X_t(\omega) = \begin{cases} \lim_n X_t^n(\omega) & \text{if } t \in T \text{ and } \omega \in A, \text{ and} \\ 0 & \text{if } t \in T \text{ and } \omega \notin A. \end{cases}$$

Then, in view of the uniform convergence of the paths, condition (a′) implies that $X_0 = 0$ and that all the paths of $\{X_t\}_{t\in T}$ are continuous. Conditions (b′) and (d′) imply that if $t_0, t_1, \ldots, t_k$ are dyadic rationals such that $t_0 < t_1 < \cdots < t_k$, then the increments $X_{t_i} - X_{t_{i-1}}$, $i = 1, \ldots, k$, are independent, with $X_{t_i} - X_{t_{i-1}}$ having distribution $N(0, t_i - t_{i-1})$. We need to extend this to the case where the t_i are not necessarily dyadic rationals.

So suppose that t_i, $i = 0, \ldots, k$, are elements of $[0,1]$ such that $t_0 < t_1 < \cdots < t_k$. Let us approximate these values by choosing sequences $\{t_{i,n}\}_n$, $i = 0, \ldots, k$, of dyadic rationals in $[0,1]$ such that $t_i = \lim_n t_{i,n}$ holds for all i and $t_{i-1,n} < t_{i,n}$ holds for all i and n. Then for each n the increments $X_{t_{i,n}} - X_{t_{i-1,n}}$, $i = 1, \ldots, k$, are independent, with $X_{t_{i,n}} - X_{t_{i-1,n}}$ having distribution $N(0, t_{i,n} - t_{i-1,n})$. The increments $X_{t_{i,n}} - X_{t_{i-1,n}}$ converge pointwise (and so[14] in distribution) to the increments $X_{t_i} - X_{t_{i-1}}$, and so it follows that the increments $X_{t_i} - X_{t_{i-1}}$, $i = 1, \ldots, k$, are independent (see Corollary 10.3.12), with $X_{t_i} - X_{t_{i-1}}$ having distribution $N(0, t_i - t_{i-1})$. This will complete the proof, as soon as we construct the processes $\{X_t^n\}_{t\in T}$, $n = 0, \ldots$.

We turn to the construction of processes $\{X_t^n\}_{t\in T}$, $n = 0, \ldots$ satisfying conditions (a′)–(d′). Recall that we have a sequence $\{Z_n\}_{n=0}^{\infty}$ of independent normal random variables, each with mean 0 and variance 1. We define the process $\{X_t^0\}_{t\in T}$ by letting $X_t^0(\omega) = tZ_0(\omega)$ hold for each ω and each t. This process certainly satisfies conditions (a′) and (b′) above.

Given the process $\{X_t^{n-1}\}_{t\in T}$, we form the process $\{X_t^n\}_{t\in T}$ as follows. For each t of the form $i/2^{n-1}$ we let $X_t^n = X_t^{n-1}$. For each t of the form $(2i+1)/2^n$, $i = 0, \ldots, 2^{n-1} - 1$, we let

$$X_t^n = X_t^{n-1} + 2^{-(n+1)/2} Z_{2^{n-1}+i}.$$

[14] Use the definition of convergence in distribution, together with the dominated convergence theorem.

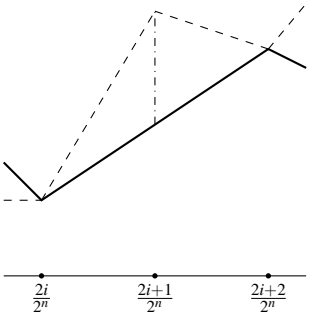

Fig. 10.1 Constructing X^n from X^{n-1}. Solid line: path of X^{n-1}. Dashed line: path of X^n. Vertical line: $2^{-(n+1)/2}Z_{2^{n-1}+i}$

Then we use straight line segments to interpolate between the points $(t, X_t^n(\omega))$, for which t has the form $i/2^n$ for some i. See Fig. 10.1. (The choice of $Z_{2^{n-1}+i}$ from the sequence of Z's is made so that the new Z's used in the construction of $\{X_t^n\}_{t\in T}$ are all distinct from those used earlier—that is, from those used in the construction of $\{X_t^k\}_{t\in T}$, where $k < n$. The coefficient of $Z_{2^{n-1}+i}$ will turn out to be what is needed to make the increments of $\{X_t^n\}_{t\in T}$ be independent and have the required distributions.) To simplify the notation a bit, let us denote $i/2^n$ by t_i, for $i = 0, \ldots, 2^n$. Then the increment $X_{t_{2i+1}}^n - X_{t_{2i}}^n$ is given by

$$\begin{aligned} X_{t_{2i+1}}^n - X_{t_{2i}}^n &= X_{t_{2i+1}}^{n-1} + 2^{-(n+1)/2}Z_{2^{n-1}+i} - X_{t_{2i}}^{n-1} \\ &= (1/2)(X_{t_{2i}}^{n-1} + X_{t_{2i+2}}^{n-1}) + 2^{-(n+1)/2}Z_{2^{n-1}+i} - X_{t_{2i}}^{n-1} \\ &= (1/2)(X_{t_{2i+2}}^{n-1} - X_{t_{2i}}^{n-1}) + 2^{-(n+1)/2}Z_{2^{n-1}+i}. \end{aligned}$$

A similar calculation shows that

$$X_{t_{2i+2}}^n - X_{t_{2i+1}}^n = (1/2)(X_{t_{2i+2}}^{n-1} - X_{t_{2i}}^{n-1}) - 2^{-(n+1)/2}Z_{2^{n-1}+i}.$$

The variables $(1/2)(X_{t_{2i+2}}^{n-1} - X_{t_{2i}}^{n-1})$ and $2^{-(n+1)/2}Z_{2^{n-1}+i}$ are independent, with each having distribution $N(0, 1/2^{n+1})$, from which it follows that the increments $X_{t_{2i+1}}^n - X_{t_{2i}}^n$ and $X_{t_{2i+2}}^n - X_{t_{2i+1}}^n$ both have distribution $N(0, 1/2^n)$. Finally, if one calculates the characteristic function of the joint distribution of the increments $X_{t_{i+1}}^n - X_{t_i}^n$, one obtains the product of the characteristic functions of normal variables with mean 0 and variance $1/2^n$, and the independence of the increments follows. With this we have verified conditions (a′), (b′), and (d′).

We turn to condition (c′), the almost sure uniform convergence of the sequence $\{X_t^n(\omega)\}$. Suppose that we can find a sequence $\{\varepsilon_n\}$ of positive numbers such that $\sum_n \varepsilon_n < +\infty$ and $\sum_n P(A_n) < +\infty$, where A_n is defined by

$$A_n = \{\sup_t |X_t^n - X_t^{n-1}| > \varepsilon_n\}.$$

Then the Borel–Cantelli lemma says that $P(\{A_n \text{ i.o.}\}) = 0$; since if $\omega \notin \{A_n \text{ i.o.}\}$, then $\sup_t |X_t^n(\omega) - X_t^{n-1}(\omega)| \le \varepsilon_n$ holds for all large n, the almost sure uniform convergence of the sequence $\{t \mapsto X_t^n(\omega)\}_{n=1}^{\infty}$ will follow from the condition $\sum_n \varepsilon_n < +\infty$.

We still need to construct the sequence $\{\varepsilon_n\}$. In view of the way $\{X_t^n\}_{t\in T}$ was constructed from $\{X_t^{n-1}\}_{t\in T}$, we have

$$\begin{aligned} P(A_n) &= P(\{\sup_t |X_t^n - X_t^{n-1}| > \varepsilon_n\}) \\ &= P(\max_{0\le i<2^{n-1}} |2^{-(n+1)/2} Z_{2^{n-1}+i}| > \varepsilon_n) \\ &\le \sum_{i=0}^{2^{n-1}-1} P(|2^{-(n+1)/2} Z_{2^{n-1}+i}| > \varepsilon_n) \\ &= 2^{n-1} P(|Z_{2^{n-1}}| > 2^{(n+1)/2}\varepsilon_n). \end{aligned}$$

Since $Z_{2^{n-1}}$ has a normal distribution with mean 0 and variance 1, it follows from Lemma 10.1.6 that

$$P(A_n) \le 2^{n-1} \frac{2}{\sqrt{2\pi} 2^{(n+1)/2}\varepsilon_n} e^{-(1/2)2^{n+1}\varepsilon_n^2} = \frac{2^{n/2-1}}{\sqrt{\pi}\varepsilon_n} e^{-2^n\varepsilon_n^2}.$$

If, for example, we let ε_n be $2^{-n/4}$, then $\sum_n \varepsilon_n < +\infty$ and $\sum_n P(A_n) < +\infty$, and the proof is complete. □

Corollary 10.5.2. *A one-dimensional Brownian motion with parameter set* $[0, +\infty)$ *exists.*

Proof. We will use a sequence $\{X_t^{(n)}\}_{t\in[0,1]}$, $n = 1, 2, \ldots$, of independent Brownian motion processes, which we can construct as follows. According to Corollary 10.1.16 there exists a sequence $\{Z_n\}$ of independent normal random variables, each with mean 0 and variance 1. Using ideas from the proof of Corollary 10.1.14, we can divide the sequence $\{Z_n\}$ into a sequence of sequences $\{Z'_{m,n}\}_m$, $n = 1, 2, \ldots$. Finally, for each n, the construction in Theorem 10.5.1 can be applied to the sequence $\{Z'_{m,n}\}_m$ to produce the process $\{X_t^{(n)}\}_{t\in[0,1]}$; the independence of these processes follows from the independence of the sequences $\{Z'_{m,n}\}_m$, $n = 1, 2, \ldots$.

Next we define a process $\{X_t\}_{t\in[0,+\infty)}$ by splicing together the paths of the processes $\{X_t^{(n)}\}_{t\in[0,1]}$—that is, by letting $X_t(\omega) = X_t^{(1)}(\omega)$ if $t \le 1$, letting $X_t(\omega) = X_1^{(1)}(\omega) + X_{t-1}^{(2)}(\omega)$ if $1 < t \le 2, \ldots$. More precisely, we define X_t recursively by

$$X_t(\omega) = \begin{cases} X_t^{(1)}(\omega) & \text{if } 0 \le t \le 1, \text{ and} \\ X_{n-1}(\omega) + X_{t-(n-1)}^{(n)}(\omega) & \text{if } n > 1 \text{ and } n-1 < t \le n. \end{cases}$$

It is clear that the paths of $\{X_t\}_{t\in[0,+\infty)}$ are continuous and that $X_0 = 0$. Now suppose that we have a sequence $t_0, t_1, \ldots, t_m$ in $[0,+\infty)$ such that $t_{i-1} < t_i$, $i = 1, \ldots, m$. Add to this sequence those integers between t_0 and t_m that are not in the original sequence, forming a new sequence $s_0, s_1, \ldots, s_n$ such that $s_{i-1} < s_i$, $i = 1, \ldots, n$. Since $\{X_t^{(n)}\}_{t\in[0,1]}$, $n = 1, 2, \ldots$, is a collection of independent Brownian motions, the increments $X_{s_i} - X_{s_{i-1}}$, $i = 1, \ldots, n$, are independent normal variables with mean 0 and the appropriate variances. It follows that the increments $X_{t_i} - X_{t_{i-1}}$, $i = 1, \ldots, m$, are independent and have the required distributions. □

Here is an interesting fact about the paths of a Brownian motion process.

Theorem 10.5.3. *Almost all the paths of a one-dimensional Brownian motion are nowhere differentiable. More precisely, let $T = [0,1]$ and let $\{X_t\}_{t\in T}$ be a one-dimensional Brownian motion on the probability space $(\Omega, \mathscr{A}, P)$. Then there is a set A in $\mathscr{A}$ such that $P(A) = 0$ and such that for each ω outside A the path $t \mapsto X_t(\omega)$ is nowhere differentiable.*

Proof. Let K be a positive integer, which we hold fixed for the moment. We will construct a sequence $\{B_n\}$ of $\mathscr{A}$-measurable subsets of Ω such that

(a) $\lim_n P(B_n) = 0$, and
(b) if ω is a element of Ω such that the path $t \mapsto X_t(\omega)$ is differentiable at some t_0 in $[0,1]$, with $|X'_{t_0}(\omega)| < K$, then ω belongs to B_n for all large n.

Suppose we have constructed such a sequence $\{B_n\}$. Let A_K be $\cup_m \cap_{n\geq m} B_n$, the set of points ω such that $\omega \in B_n$ holds for all large n. Then $P(\cap_{n\geq m} B_n) \leq \lim_n P(B_n) = 0$ holds for all m, and so $P(A_K) = 0$. Now suppose that we let K vary through the positive integers, and we define A by $A = \cup_{K=1}^{\infty} A_K$. Then A has P-measure 0, and it follows from condition (b) that A contains every ω for which the path $t \mapsto X_t(\omega)$ is differentiable at one or more points; in other words, A is as described in the statement of the theorem.

Now we turn to our remaining task, the construction of a sequence[15] $\{B_n\}$ of sets satisfying conditions (a) and (b) above. We once again consider K to be fixed; we do so through the end of the proof. For each n, where $n \geq 3$, we define sets $C_{n,k}$, $k = 1, \ldots, n$ by

$$C_{n,k} = \left\{ \omega : |X_{k/n}(\omega) - X_{(k-1)/n}(\omega)| < \frac{3K}{n} \right\},$$

and then we define sets $D_{n,k}$, $k = 2, \ldots, n-1$ by

$$D_{n,k} = C_{n,k-1} \cap C_{n,k} \cap C_{n,k+1}.$$

Finally, we define sets B_n by $B_n = C_{n,1} \cup C_{n,n} \cup (\cup_{k=2}^{n-1} D_{n,k})$. We will show that the sets B_n satisfy (a) and (b).

[15] Our sequence will start with n equal to 3, rather than equal to 0 or 1.

We begin our verification of condition (a) by estimating the probabilities of the sets $C_{n,k}$. Since the difference $X_{k/n} - X_{(k-1)/n}$ is normal with mean 0 and variance $1/n$, it has the same distribution as the variable $Z/\sqrt{n}$, where Z is a normal variable with mean 0 and variance 1. Thus

$$P(C_{n,k}) = P\left(|X_{k/n} - X_{(k-1)/n}| < \frac{3K}{n}\right) = P\left(|Z| < \frac{3K}{\sqrt{n}}\right)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\frac{3K}{\sqrt{n}}}^{\frac{3K}{\sqrt{n}}} e^{-x^2/2}\, dx < \frac{K_1}{\sqrt{n}},$$

where K_1 is the constant $6K/\sqrt{2\pi}$. The independence of the events $C_{n,k}$, $k = 1, \ldots, n$, implies that

$$P(D_{n,k}) = P(C_{n,k-1})P(C_{n,k})P(C_{n,k+1}) < {K_1}^3/n^{3/2}.$$

Since $B_n = C_{n,1} \cup C_{n,n} \cup (\cup_{k=2}^{n-1} D_{n,k})$, we have $P(B_n) < 2K_1/\sqrt{n} + (n-2){K_1}^3/n^{3/2}$, and $\lim_n P(B_n) = 0$ follows. Thus condition (a) holds.

We turn to condition (b). Suppose that $t \mapsto X_t(\omega)$ is differentiable at the point t_0, and that $|X'_{t_0}(\omega)| < K$. Let n be large enough that

$$|X_t(\omega) - X_{t_0}(\omega)| < K|t - t_0| \tag{1}$$

holds when $|t - t_0| \le 2/n$. It follows that if $t_0 \in [\frac{k-1}{n}, \frac{k}{n}]$, then

$$|X_{k/n}(\omega) - X_{(k-1)/n}(\omega)| < K/n,$$

while if t_0 lies in an interval of length $1/n$ adjacent to the interval $[\frac{k-1}{n}, \frac{k}{n}]$, then

$$|X_{k/n}(\omega) - X_{(k-1)/n}(\omega)| \le |X_{k/n}(\omega) - X_{t_0}(\omega)| + |X_{t_0}(\omega) - X_{(k-1)/n}(\omega)|$$
$$< K/n + 2K/n = 3K/n.$$

Now suppose that k is such that $t_0 \in [\frac{k-1}{n}, \frac{k}{n}]$. The estimates we have just made show that $\omega \in C_{n,1} \cup C_{n,n}$ if k is 1 or n and that $\omega \in D_{n,k}$ otherwise. In any case, $\omega \in B_n$, and the verification of condition (b) is complete. □

Exercises

1. Suppose that we have a stochastic process $\{X_t\}$ with index set $[0,1] \cap \mathbb{Q}$ that satisfies properties (a) and (b) in the definition of Brownian motion (where the values t_i are restricted to lie in $[0,1] \cap \mathbb{Q}$). In this exercise we prove that almost all the paths of this process are uniformly continuous on $[0,1] \cap \mathbb{Q}$. In the following

exercise we use this to give another construction of a Brownian motion process on $[0,1]$.

(a) Show that if a and b are rational numbers that satisfy $0 \le a < b \le 1$ and if C is a positive constant, then

$$P(\sup\{X_t - X_a : t \in [a,b] \cap \mathbb{Q}\} > C) \le 2P(X_b - X_a > C). \tag{2}$$

(Hint: First suppose that $a \le t_1 < t_2 < \cdots < t_k \le b$ and let A_i be the event that i is the smallest value of j for which $X_{t_j} - X_a > C$. Check that

$$\begin{aligned} P(A_i) &= P(A_i \cap \{X_b - X_{t_i} \ge 0\}) + P(A_i \cap \{X_b - X_{t_i} < 0\}) \\ &\le 2P(A_i \cap \{X_b - X_a > C\}), \end{aligned}$$

and then use this estimate to prove the analogue of (2) in which the supremum is taken as t ranges over $\{t_1, t_2, \ldots, t_n\}$. Finally, take limits as more and more points from $[a,b] \cap \mathbb{Q}$ are considered in the supremum.)

(b) For each positive δ define $v(\delta)$ by

$$v(\delta) = \sup\{|X_t - X_s| : s,t \in [0,1] \cap \mathbb{Q} \text{ and } |t-s| < \delta\}.$$

Use part (a), together with Lemma 10.1.6, to show that there exist sequences $\{\varepsilon_n\}$ and $\{\delta_n\}$ of positive numbers such that $\lim_n \varepsilon_n = \lim_n \delta_n = 0$ and

$$\sum_i P(v(\delta_n) > \varepsilon_n) < +\infty;$$

from this derive the almost sure uniform continuity of the paths.

2. In Exercise 10.6.4 we will construct a stochastic process $\{X_t\}$ with index set $[0,1] \cap \mathbb{Q}$ that satisfies properties (a) and (b) in the definition of Brownian motion. Given that result, use Exercise 1 to give a proof of the existence of Brownian motion on $[0,1]$ that is quite different from the proof in the text.
3. Let $T = [0, +\infty)$, let $(\Omega, \mathscr{A}, P)$ be a probability space, and let $\{X_t\}_{t \in T}$ be a Brownian motion process on $(\Omega, \mathscr{A}, P)$. Define a filtration $\{\mathscr{F}_t\}_{t \in T}$ by letting $\mathscr{F}_t = \sigma(\{X_s : s \le t\})$ hold for each t in T.

(a) Let a be a real number. Show that the function $\tau\colon \Omega \to [0, +\infty]$ defined by $\tau(\omega) = \inf\{t : X_t(\omega) = a\}$ is a stopping time.

(b) Suppose τ is a stopping time. Show that if n is a positive integer, then

$$\tau_n(\omega) = \inf\{i/2^n : \tau(\omega) \le i/2^n\}$$

defines a stopping time (of course, $\tau_n(\omega) = +\infty$ if $\tau(\omega) = +\infty$).

(c) Show that if τ is a stopping time, then X_τ is $\mathscr{F}_\tau$-measurable.

4. Let $T = [0, +\infty)$ and let $\{X_t\}_{t \in T}$ be a Brownian motion process.

(a) Fix a value t_0 such that $0 < t_0 < +\infty$ and define a process $\{Y_t\}_{t\in T}$ by $Y_t = X_{t+t_0} - X_{t_0}$ for t in T. Show that $\{Y_t\}_{t\in T}$ is a Brownian motion and that it is independent of $\mathscr{F}_{t_0}$ (in other words, the σ-algebras $\sigma(Y_t, t \in T)$ and $\mathscr{F}_t$ are independent).

(b) Suppose that τ is a stopping time that is finite almost surely, and define a process $\{Y_t\}_{t\in T}$ by

$$Y_t(\omega) = \begin{cases} X_{t+\tau(\omega)}(\omega) - X_{\tau(\omega)}(\omega) & \text{if } \tau(\omega) < +\infty, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Show that if the stopping time τ has only finitely many values, then $\{Y_t\}_{t\in T}$ is a Brownian motion that is independent of $\mathscr{F}_\tau$.

(c) Show that the assumption that τ has only finitely many values can be removed from part (b). (Hint: See Exercise 3.)

10.6 Construction of Probability Measures

This section contains two constructions of possibly infinite families of random variables with specified distributions. The first construction gives sequences of independent random variables, while the second gives families of not necessarily independent random variables.

Let us recall the methods we have been using to construct sequences of independent real-valued random variables. In simple cases, where we need only finitely many independent random variables, say with distributions $\mu_1, \mu_2, \ldots, \mu_d$, we saw that we can take the product measure $\mu_1 \times \cdots \times \mu_d$ on $\mathbb{R}^d$ and then let the random variables be the coordinate functions on $\mathbb{R}^d$. On the other hand, to construct an infinite sequence of independent real-valued random variables, we used a perhaps awkward-seeming ad hoc construction based on the binary expansion of numbers in the unit interval, together with a kind of inverse for distribution functions of probability measures (see the end of Sect. 10.1).

Here we will look at the use of product spaces to construct infinite families of random variables. Note that the random variables we construct do not need to be real valued—in our first construction, they can have values in arbitrary measurable spaces, while in our second construction, they can have values in rather general, but not arbitrary, spaces.

We begin by defining the measurable spaces on which we will construct families of random variables. Let I be an index set, and let $\{(\Omega_i, \mathscr{A}_i)\}_{i\in I}$ be an indexed family of measurable spaces. (In typical situations the measurable spaces $(\Omega_i, \mathscr{A}_i)$ will be equal to one another.) The *product* of these measurable spaces is the measurable space $(\Omega, \mathscr{A})$ defined as follows: The underlying set Ω is the product $\prod_i \Omega_i$ of the sets $\{\Omega_i\}_i$; that is, Ω is the set of all functions $\omega\colon I \to \cup_i \Omega_i$ such that $\omega(i) \in \Omega_i$ for each i in I. For each i we define the coordinate function $X_i\colon \Omega \to \Omega_i$ by

$X_i(\omega) = \omega(i)$. Finally, we let $\mathscr{A}$ be the smallest σ-algebra on Ω that makes each X_i measurable with respect to $\mathscr{A}$ and $\mathscr{A}_i$. Equivalently, we can let $\mathscr{A}$ be the σ-algebra on Ω generated by the collection of all sets that have the form

$$\{\omega \in \Omega : \omega(i) \in A_i \text{ holds for each } i \text{ in } I_0\}$$

for some finite subset I_0 of I and some sets A_i that satisfy $A_i \in \mathscr{A}_i$ for each i in I_0.

Let us turn to the construction of sequences of independent random variables.

Proposition 10.6.1. *Let $\{(\Omega_i, \mathscr{A}_i, P_i)\}_{i\in\mathbb{N}}$ be a family of probability spaces indexed by the set $\mathbb{N}$ of positive integers, let $(\Omega, \mathscr{A})$ be the product of the measurable spaces $\{(\Omega_i, \mathscr{A}_i)\}_{i\in\mathbb{N}}$, and for each i in $\mathbb{N}$ let X_i be the coordinate projection from Ω to Ω_i. Then there is a unique probability measure P on $(\Omega, \mathscr{A})$ such that*

(a) *for each i the distribution of X_i is P_i, and*
(b) *the random variables $\{X_i\}_{i\in\mathbb{N}}$ are independent.*

Proof. What we need here is a product measure with infinitely many factors. In particular, we need a measure P on $(\Omega, \mathscr{A})$ such that for each n and each choice of sets A_i in $\mathscr{A}_i$, $i = 1, \ldots, n$, we have

$$P(A) = P_1(A_1)\,P_2(A_2)\cdots P_n(A_n),$$

where A is the subset

$$A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \cdots \tag{1}$$

of Ω—that is, where A consists of those sequences $\{x_i\}_1^\infty$ in Ω such that $x_i \in A_i$ holds for $i = 1, \ldots, n$.

The results in Chap. 5 give us a start on the construction of such measures. Namely for each n those results give us a product measure $P_1 \times \cdots \times P_n$ on the measurable space $(\prod_1^n \Omega_i, \prod_1^n \mathscr{A}_i)$. For each n let proj_n be the projection of the infinite product Ω onto $\prod_1^n \Omega_i$, that is, the function that takes an infinite sequence to the sequence of its first n components. Let $\mathscr{A}^{(1)}$ be the collection of subsets of Ω defined[16] by

$$\mathscr{A}^{(1)} = \bigcup_n \text{proj}_n^{-1}(\prod_1^n \mathscr{A}_i).$$

Since $\{\text{proj}_n^{-1}(\prod_1^n \mathscr{A}_i)\}_{n=1}^\infty$ is an increasing sequence of σ-algebras on Ω, it follows that $\mathscr{A}^{(1)}$ is an algebra of sets. Furthermore $\mathscr{A} = \sigma(\mathscr{A}^{(1)})$. We need to transfer our finite-dimensional product measures to $\mathscr{A}^{(1)}$. For that, define a function P on $\mathscr{A}^{(1)}$ by letting

$$P(\text{proj}_n^{-1}(A)) = (P_1 \times \cdots \times P_n)(A)$$

[16]Note that if X and Y are sets, if f is a function from X to Y, and if $\mathscr{C}$ is a family of subsets of Y, then $f^{-1}(\mathscr{C}) = \{f^{-1}(C) : C \in \mathscr{C}\}$.

hold for each n and each A in $\prod_1^n \mathscr{A}_i$ (the reader should check that P is well defined). Certainly P has the necessary value on each rectangular set of the form given in (1). Furthermore P is countably additive on each $\text{proj}_n^{-1}(\prod_1^n \mathscr{A}_i)$ and so is at least finitely additive on $\mathscr{A}^{(1)}$. If we show that P is countably additive on $\mathscr{A}^{(1)}$, then it will have a countably additive extension to $\mathscr{A}$ (see Exercise 1.3.5) and the proof of existence will be complete.

We need a bit of notation for the proof of countable additivity. For each n we want the analogue of Ω, $\mathscr{A}^{(1)}$, and P, but with the products starting with $(\Omega_n, \mathscr{A}_n)$ and P_n, rather than with $(\Omega_1, \mathscr{A}_1)$ and P_1. Let us use the notation $\Omega^{(n)}$, $\mathscr{A}^{(n)}$, and $P^{(n)}$ for such sets,[17] algebras, and finitely additive probabilities. Note that $\Omega^{(1)} = \Omega$, $P^{(1)} = P$, and $\mathscr{A}^{(1)}$ is the algebra discussed above. Note also that if A is a set in $\mathscr{A}^{(n)}$, then for each x in Ω_n the section A_x belongs to $\mathscr{A}^{(n+1)}$. Finally, let us introduce the following temporary notation for sections of sets. Instead of writing A_x we will write $A(x)$, and instead of writing $(A_{x_1})_{x_2}$ we will write $A(x_1, x_2)$. Continuing in this way gives a reasonable way to express the result of many iterations of the operation of taking a section of a set.

We prove the countable additivity of P by showing that if $\{A_j\}$ is a decreasing sequence of sets in $\mathscr{A}^{(1)}$ such that $\cap_j A_j = \varnothing$, then $\lim_j P(A_j) = 0$.[18] We do this by considering the contrapositive and showing that if $\{A_j\}$ is a decreasing sequence of sets in $\mathscr{A}^{(1)}$ such that $\lim_j P(A_j) > 0$, then $\cap_j A_j \neq \varnothing$. So let us fix a decreasing sequence $\{A_j\}$ and a positive number ε such that $P(A_j) \geq \varepsilon$ holds for all j. We will show that $\cap_j A_j \neq \varnothing$ by constructing an element of $\cap_j A_j$. Suppose that A_j is a member of the sequence $\{A_j\}$. Then there is a positive integer k and a set B_j in $\prod_1^k \mathscr{A}_i$ such that $A_j = \text{proj}_k^{-1}(B_j)$. We have (see Theorem 5.1.4)

$$(P_1 \times \cdots \times P_k)(B_j) = \int_{\Omega_1} (P_2 \times \cdots \times P_k)(B_j(x_1))\, P_1(dx_1),$$

which translates into

$$P(A_j) = \int_{\Omega_1} P^{(2)}(A_j(x_1))\, P_1(dx_1).$$

Since $\{A_j\}_j$ is a decreasing sequence of sets, $\{P^{(2)}(A_j(x_1))\}_j$ is (for each choice of x_1 in Ω_1) a decreasing sequence of numbers, and we can define a function $f_1 : \Omega_1 \to \mathbb{R}$ by $f_1(x_1) = \lim_j P^{(2)}(A_j(x_1))$. The function f_1 is measurable, and it follows from the dominated convergence theorem that

$$\int_{\Omega_1} f_1(x_1)\, P_1(dx_1) = \lim_j \int_{\Omega_1} P^{(2)}(A_j(x_1))\, P_1(dx_1) = \lim_j P(A_j) \geq \varepsilon.$$

[17] Be careful to note that $\Omega^{(n)}$ is a product space, while Ω_n is one of (in fact, the first of) its factors.

[18] See Proposition 1.2.6, whose proof can easily be modified so as to apply to finitely additive measures *on algebras*.

Since P_1 has total mass 1, there must be an element x_1 of Ω_1 such that $f_1(x_1) \geq \varepsilon$ and hence such that $P^{(2)}(A_j(x_1)) \geq \varepsilon$ holds for all j; fix such a value x_1. We can apply the same argument to the sequence $\{A_j(x_1)\}_j$, producing an element x_2 of Ω_2 such that $P^{(3)}(A_j(x_1,x_2)) \geq \varepsilon$ holds for each j. By repeating this argument over and over, we produce a sequence $\{x_n\}$ such that $P^{(n+1)}(A_j(x_1,\ldots,x_n)) \geq \varepsilon$ holds for all j and n.

To complete our proof that P is countably additive on $\mathscr{A}^{(1)}$, we need to show that $\cap_j A_j \neq \varnothing$. We do this by verifying that the sequence $\{x_n\}$ constructed above belongs to $\cap_j A_j$. So fix a set A_j in $\{A_j\}$. Then there is a positive integer k and a set B_j in $\prod_1^k \mathscr{A}_i$ such that $A_j = \text{proj}_k^{-1}(B_j)$. Note that, because of this representation of A_j, the section $A_j(x_1,\ldots,x_k)$ is equal to either $\Omega^{(k+1)}$ or $\varnothing$, depending on whether $(x_1,\ldots,x_k)$ belongs to B_j or not. However, we know that $A_j(x_1,\ldots,x_k)$ is not empty (since $P^{(k+1)}(A_j(x_1,\ldots,x_k)) \geq \varepsilon$). Thus, $A_j(x_1,\ldots,x_k) = \Omega^{(k+1)}$, and *every* continuation of the finite sequence $x_1, \ldots, x_k$ belongs to A_j; in particular $\{x_n\} \in A_j$. Since this argument works for every j, we have $\{x_n\} \in \cap A_j$, and the construction of our product measure is complete.

We turn to the uniqueness of P. The collection of sets of the form (1) (where $A_i \in \mathscr{A}_i$ holds for each i) is a π-system that generates $\mathscr{A}$, and so the uniqueness of P follows from Corollary 1.6.3. □

See Exercise 2 for an extension of Proposition 10.6.1 to the case of uncountably many random variables.

Now we turn to the construction of families of random variables that are not necessarily independent. For the construction of such families we will once again build a suitable measure on an infinite product space. This time, however, the measure we construct will not be a product measure.

As before, let I be an index set and let $\{(\Omega_i, \mathscr{A}_i)\}_{i\in I}$ be an indexed family of measurable spaces. Let $(\Omega, \mathscr{A})$ and $\{X_i\}_{i\in I}$ be the measurable space and coordinate functions constructed at the beginning of this section. We need to look at how to describe the dependence between our random variables. To get an idea of what to do, let us temporarily assume that we already have a probability P on $(\Omega, \mathscr{A})$. We will get a consistency condition that the joint distributions of finite collections of the random variables $\{X_i\}$ must satisfy; then we will use this consistency condition as one of the hypotheses in our existence theorem (Theorem 10.6.2).

Let $\mathscr{I}$ be the collection of all nonempty finite subsets of I. For each I_0 in $\mathscr{I}$ consider the finite product $(\prod_{i\in I_0}\Omega_i, \prod_{i\in I_0}\mathscr{A}_i)$. Let us call this product $(\Omega_{I_0}, \mathscr{A}_{I_0})$. For each I_0 let $X_{I_0} : \Omega \to \Omega_{I_0}$ be the projection of Ω onto Ω_{I_0}. So in set-theoretic terms, $X_{I_0}(\omega)$ is the restriction of the function ω to the subset I_0 of its domain. It is easy to check that for each I_0 the function X_{I_0} is measurable with respect to $\mathscr{A}$ and $\mathscr{A}_{I_0}$. Let P_{I_0} be the distribution of X_{I_0} (in other words, let P_{I_0} be the joint distribution of the random variables X_i, $i \in I_0$); thus $P_{I_0}(A) = P(X_{I_0}^{-1}(A))$ holds for each A in $\mathscr{A}_{I_0}$.

We need to look at how these distributions on finite products are related to one another. So suppose that I_1 and I_2 belong to $\mathscr{I}$ and satisfy $I_2 \subseteq I_1$, and let $\text{proj}_{I_2,I_1} : \Omega_{I_1} \to \Omega_{I_2}$ be the projection of Ω_{I_1} onto Ω_{I_2}. Certainly proj_{I_2,I_1} is

measurable and $X_{I_2} = \text{proj}_{I_2,I_1} \circ X_{I_1}$; thus $P(X_{I_2}^{-1}(A)) = P(X_{I_1}^{-1}(\text{proj}_{I_2,I_1}^{-1}(A)))$ holds for each A in $\mathscr{A}_{I_2}$. That is, the distributions on the finite product spaces satisfy the condition

$$P_{I_2} = P_{I_1} \text{proj}_{I_2,I_1}^{-1} \text{ for all } I_1, I_2 \text{ in } \mathscr{I} \text{ such that } I_2 \subseteq I_1. \tag{2}$$

This is the consistency condition that will be one of the hypotheses in the following theorem.

The upcoming theorem would not hold if the spaces $(\Omega_i, \mathscr{A}_i)$ were allowed to be completely arbitrary (see Exercise 5). To get around that difficulty, we will assume that for each i there is a compact metric space K_i such that $(\Omega_i, \mathscr{A}_i)$ is Borel isomorphic to $(K_i, \mathscr{B}(K_i))$; in other words, there must be a bijection $f_i \colon \Omega_i \to K_i$ such that f_i and f_i^{-1} are both measurable. Such measurable spaces are called *standard*.[19] One can check (see Exercise 1) that $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is isomorphic to $([0,1], \mathscr{B}([0,1]))$ and hence that $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is standard; from this one can conclude that $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ is also standard.

Theorem 10.6.2 (Kolmogorov Consistency Theorem). *Let I be a nonempty set, let $\{(\Omega_i, \mathscr{A}_i)\}_{i \in I}$ be an indexed family of measurable spaces, and let $\mathscr{I}$ be the collection of all nonempty finite subsets of I. As in the discussion above, define product measurable spaces $(\Omega, \mathscr{A})$ and $\{(\Omega_{I_0}, \mathscr{A}_{I_0})\}_{I_0 \in \mathscr{I}}$, plus projections $X_{I_0} \colon \Omega \to \Omega_{I_0}$ and $\text{proj}_{I_2,I_1} \colon \Omega_{I_1} \to \Omega_{I_2}$, where $I_0, I_1, I_2 \in \mathscr{I}$ and $I_2 \subseteq I_1$. Let $\{P_{I_0}\}_{I_0 \in \mathscr{I}}$ be an indexed family of probability measures on the spaces $\{(\Omega_{I_0}, \mathscr{A}_{I_0})\}_{I_0 \in \mathscr{I}}$. If*

(a) *the measurable spaces $\{(\Omega_i, \mathscr{A}_i)\}_{i \in I}$ are all standard, and*
(b) *the measures $\{P_{I_0}\}_{I_0 \in \mathscr{I}}$ are consistent, in the sense that they satisfy condition (2),*

then there is a unique probability measure P on $(\Omega, \mathscr{A})$ such that for each I_0 in $\mathscr{I}$ the distribution of X_{I_0} is P_{I_0}.

Proof. The hypothesis that the spaces $\{(\Omega_i, \mathscr{A}_i)\}_{i \in I}$ are standard implies that for each i there is a compact metrizable topology on Ω_i for which $\mathscr{B}(\Omega_i) = \mathscr{A}_i$. Fix such a topology for each i. It follows from Tychonoff's theorem (Theorem D.20) and Proposition 7.1.13 that the product topology on Ω is compact Hausdorff and that for each I_0 the product topology on Ω_{I_0} is compact and metrizable; furthermore, $\mathscr{B}(\Omega_{I_0}) = \mathscr{A}_{I_0}$ holds for each I_0 in $\mathscr{I}$ (see Proposition 7.6.2). We will construct a suitable positive linear functional L on the space $C(\Omega)$ of continuous real-valued functions on Ω. The Riesz representation theorem (Theorem 7.2.8) then gives a regular Borel measure μ on Ω such that $L(f) = \int f \, d\mu$ holds for each f in $C(\Omega)$. We will see that the restriction of μ to $\mathscr{A}$ is the measure we need.

We turn to the definition of the linear functional L. We begin by defining it on the algebra of functions on Ω generated by the functions that can be written in the form $g \circ X_i$ for some i in I and some g in $C(\Omega_i)$. Let us call this algebra $C_\bullet$. Since the functions h in $C_\bullet$ are finite sums of finite products of functions of the form $g \circ X_i$,

[19] See Chap. 8, and especially Sect. 8.6, for more information about standard spaces.

each can be written in the form $h_{I_0} \circ X_{I_0}$ for some I_0 in $\mathscr{I}$ and some h_{I_0} in $C(\Omega_{I_0})$. We want to define $L(h)$ for h in $C_\bullet$ by $L(h) = \int_{\Omega_{I_0}} h_{I_0}\, dP_{I_0}$, where h and h_{I_0} are related by $h = h_{I_0} \circ X_{I_0}$. The potential problem is that a function h can in general be written in many ways, say as $h_{I_1} \circ X_{I_1}$ and as $h_{I_2} \circ X_{I_2}$, and so we need to check that $L(h)$ does not depend on how h is written.[20] Suppose that I_1 and I_2 are as in the previous sentence, and let $I_3 = I_1 \cup I_2$. The relation $h_{I_1} \circ X_{I_1} = h = h_{I_2} \circ X_{I_2}$ implies that

$$h_{I_1} \circ \mathrm{proj}_{I_1,I_3} = h_{I_2} \circ \mathrm{proj}_{I_2,I_3}\,.$$

From this and the consistency condition (2), we find

$$\begin{aligned}\int_{\Omega_{I_1}} h_{I_1}\, dP_{I_1} &= \int_{\Omega_{I_3}} h_{I_1} \circ \mathrm{proj}_{I_1,I_3}\, dP_{I_3} \\ &= \int_{\Omega_{I_3}} h_{I_2} \circ \mathrm{proj}_{I_2,I_3}\, dP_{I_3} = \int_{\Omega_{I_2}} h_{I_2}\, dP_{I_2},\end{aligned}$$

and it follows that L is well defined on $C_\bullet$. The Stone–Weierstrass theorem (Theorem D.22) implies that $C_\bullet$ is uniformly dense in $C(\Omega)$. Thus we can extend L from $C_\bullet$ to $C(\Omega)$. It is easy to check that the extended L is positive and linear. Thus the Riesz representation theorem gives a regular Borel measure μ on Ω such that $L(h) = \int h\, d\mu$ holds for each h in $C(\Omega)$. In particular, for each I_0 in $\mathscr{I}$ and each h_{I_0} in $C(\Omega_{I_0})$ we have

$$\int_{\Omega_{I_0}} h_{I_0}\, dP_{I_0} = L(h_{I_0} \circ X_{I_0}) = \int_{\Omega} h_{I_0} \circ X_{I_0}\, d\mu = \int_{\Omega_{I_0}} h_{I_0}\, d(\mu X_{I_0}^{-1}). \tag{3}$$

Let P be the restriction of μ to $\mathscr{A}$. It follows from Eq. (3) that $P_{I_0} = PX_{I_0}^{-1}$. In other words, P_{I_0} is the distribution of X_{I_0} under P. Since this is true for each I_0 in $\mathscr{I}$, we have constructed the required measure on $(\Omega, \mathscr{A})$.

We turn to the uniqueness of P. Define $\mathscr{A}'$ by $\mathscr{A}' = \cup_{I_0 \in \mathscr{I}} X_{I_0}^{-1}(\mathscr{A}_{I_0})$. Then $\mathscr{A}'$ is a π-system on Ω and $\sigma(\mathscr{A}') = \mathscr{A}$. Suppose that P' and P'' are probabilities on $\mathscr{A}$ that satisfy $P_{I_0} = P'X_{I_0}^{-1} = P''X_{I_0}^{-1}$ for each I_0 in $\mathscr{I}$. This means that P' and P'' agree on $\mathscr{A}'$, and it follows from Corollary 1.6.3 that $P' = P''$. With this the proof is complete. □

Exercises

1. Check that the measurable spaces $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and $([0,1], \mathscr{B}([0,1]))$ are isomorphic. (Hint: This is an immediate consequence of some of the results in Chap. 8.

[20] This is where we use the consistency condition (2).

A more elementary proof is possible: start with a homeomorphism of $\mathbb{R}$ onto the open interval $(0,1)$, and modify it on a countable set so as to get a suitable bijection from $\mathbb{R}$ onto the closed interval $[0,1]$.)

2. Show that Proposition 10.6.1 also holds for uncountable families of independent random variables (i.e., for uncountable index sets). (Hint: Suppose that the index set I is uncountable. Combine the version of Proposition 10.6.1 for countable products with the fact that the product σ-algebra on $\prod_{i\in I}\Omega_i$ is the union of the inverse images (under projection) of the product σ-algebras on the countable products $\prod_{i\in I_0}\Omega_i$, where I_0 ranges over the countable subsets of I. See Exercise 1.1.7.)
3. Let $T=[0,1]$. For each t in T let $(\Omega_t,\mathscr{A}_t)=(\mathbb{R},\mathscr{B}(\mathbb{R}))$, and let $(\Omega,\mathscr{A})$ be the product of these spaces. Show that the subset of Ω consisting of the continuous functions from T to $\mathbb{R}$ does not belong to $\mathscr{A}$.[21] (Hint: See Exercise 1.1.7.)
4. Use Theorem 10.6.2 to construct a stochastic process $\{X_t\}$ with index set $[0,1]\cap\mathbb{Q}$ that satisfies properties (a) and (b) in the definition of Brownian motion (where the values t_i are restricted to lie in $[0,1]\cap\mathbb{Q}$). (Given this result, Exercises 10.5.1 and 10.5.2 can be used to give a proof of Theorem 10.5.1 that is less technical than the one given in Sect. 10.5.)
5. Show that the conclusion of the Kolmogorov consistency theorem (Theorem 10.6.2) may fail if the assumption that the measurable spaces $(\Omega_i,\mathscr{A}_i)$ are standard is simply omitted. (Hint: Let $\{A_n\}$ be a decreasing sequence of subsets of $[0,1]$ such that $\lambda^*(A_n)=1$ holds for each n, but for which $\cap_n A_n=\varnothing$. See Exercise 1.4.7. For each n let $\Omega_n=A_n$ and let $\mathscr{A}_n$ be the trace of $\mathscr{B}(\mathbb{R})$ on A_n. Finally, for index sets I_0 of the form $\{1,2,\ldots,n\}$ define P_{I_0} on $(\Omega_{I_0},\mathscr{A}_{I_0})$ by letting it be the image of the trace of Lebesgue measure on A_n under the mapping $x\mapsto(x,x,\ldots,x)$.)
6. Assume that we modify the statement of the Kolmogorov consistency theorem (Theorem 10.6.2) by replacing the assumption that the spaces $(\Omega_i,\mathscr{A}_i)$ are standard with the assumption that each Ω_i is a universally measurable subset of some compact metric space K_i (and adding the assumption that $\mathscr{A}_i$ is the trace of $\mathscr{B}(K_i)$ on K_i). Prove that this modified version is true. (Hint: Don't work too hard—derive this modified version from the original version of Theorem 10.6.2.)

Notes

Kolmogorov was at the forefront of early work on measure-theoretic probability, as was Doob a few years later; see Kolmogorov's book on the foundations of

[21] Thus one often needs to say things like "There is a set A in $\mathscr{A}$ that has probability 1 and is such that $t\mapsto X_t(\omega)$ is continuous for each ω in A." rather than less pedantic things like "The set of all ω such that $t\mapsto X_t(\omega)$ is continuous has probability 1."

probability [72] and Doob's book on stochastic processes [38]. Dudley [40] gives detailed historical citations in his end-of-chapter notes.

See Billingsley [8], Dudley [40], Klenke [71], Lamperti [79], Walsh [124], and Williams [128] for introductions to probability that carry the ideas in this chapter much further and are at a level appropriate for people who have completed a course in measure theory.

Much more on dealing with convergence of probability measures using distances (see a remark near the start of Sect. 10.3, and see Exercise 10.3.12) can be found in Dudley [40] and Dudley [41].