# APPENDIX A

# PROBABILITY AND STOCHASTIC PROCESSES

The purpose of this chapter is to review some fundamental concepts in probability and stochastic processes, and to familiarize the reader with the notation in this book.

## A.1 RANDOM EXPERIMENTS AND PROBABILITY SPACES

The basic notion in probability theory is that of a **random experiment**: an experiment whose outcome cannot be determined in advance. Mathematically, a random experiment is modeled via a **probability space** $(\Omega, \mathcal{H}, \mathbb{P})$, where:

- $\Omega$ is the set of all possible outcomes of the experiment, called the **sample space**.
- $\mathcal{H}$ is the collection of all subsets of $\Omega$ to which a probability can be assigned; such subsets are called **events**. The collection $\mathcal{H}$ is assumed to contain $\Omega$ itself, be closed under complements ($A \in \mathcal{H} \Rightarrow A^c \in \mathcal{H}$), and be closed under countable unions ($A_1, A_2, \ldots \in \mathcal{H} \Rightarrow \cup_i A_i \in \mathcal{H}$). Such a collection is called a **$\sigma$-algebra**.
- $\mathbb{P}$ is a **probability measure**, which assigns to each event $A$ a number $\mathbb{P}(A)$ between 0 and 1, indicating the likelihood that the outcome of the random experiment lies in $A$.

Any probability measure $\mathbb{P}$ must satisfy the following **Kolmogorov axioms**:

1. $\mathbb{P}(A) \geqslant 0$ for all $A \in \mathcal{H}$.
2. $\mathbb{P}(\Omega) = 1$.
3. For any sequence $A_1, A_2, \ldots$ of disjoint (that is, nonoverlapping) events,

$$\mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i)\,. \tag{A.1}$$

The axioms ensure that the probability of any event lies between 0 and 1. An event that happens with probability 1 is called an **almost sure** (a.s.) event. The requirement (A.1) is often referred to as the **sum rule** of probability. It simply states that if an event can happen in a number of different but not simultaneous ways, the probability of that event is the sum of the probabilities of the comprising events.

### ■ EXAMPLE A.1 (Discrete Sample Space)

In many applications the sample space is **countable**, that is, $\Omega = \{a_1, a_2, \ldots\}$. In this case the easiest way to specify a probability measure $\mathbb{P}$ is to first assign a probability $p_i$ to each **elementary event** $\{a_i\}$, with $\sum_i p_i = 1$, and then to define

$$\mathbb{P}(A) = \sum_{i:a_i \in A} p_i \quad \text{for all } A \subseteq \Omega\,.$$

Here the collection of events $\mathcal{H}$ can be taken to be equal to the collection of *all* subsets of $\Omega$. The triple $(\Omega, \mathcal{H}, \mathbb{P})$ is called a **discrete probability space**.

This idea is graphically represented in Figure A.1. Each element $a_i$, represented by a black dot, is assigned a probability weight $p_i$, indicated by the size of the dot. The probability of the event $A$ is simply the sum of the weights of all the outcomes in $A$.

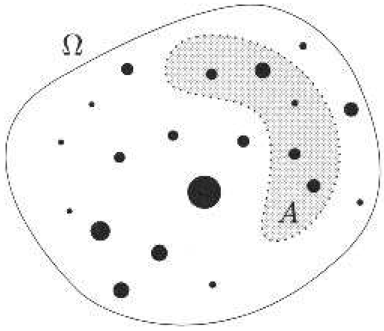

**Figure A.1** A discrete sample space.

**Remark A.1.1 (Equilikely Principle)** A special case of a discrete probability space occurs when a random experiment has finitely many and *equally likely* outcomes. In this case the probability measure is given by

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}\,, \tag{A.2}$$

where $|A|$ denotes the number of outcomes in $A$ and $|\Omega|$ the total number of outcomes. Thus, the calculation of probabilities reduces to counting. This is called the **equilikely principle**.

### A.1.1 Properties of a Probability Measure

The following properties of a probability measure follow directly from the Kolmogorov axioms. Proofs can be found, for example, in [5, 25].

1. *Complement*: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

2. *Monotonicity*: $A \subseteq B \Rightarrow \mathbb{P}(A) \leqslant \mathbb{P}(B)$.

3. *Sum rule*: $\{A_i\}$ disjoint $\Rightarrow \mathbb{P}\left(\cup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.

4. *Inclusion–exclusion*:

$$\mathbb{P}\left(\cup_i A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots .$$

   In particular, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

5. *Continuity from below*: Let $A_1, A_2, \ldots$ be an increasing sequence of events, that is, $A_1 \subseteq A_2 \subseteq \cdots \subseteq A$, with $A = \cup_n A_n$. Then, the sequence $\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots$ increases monotonely to $\mathbb{P}(A)$.

6. *Continuity from above*: Let $A_1, A_2, \ldots$ be a decreasing sequence of events, that is, $A_1 \supseteq A_2 \supseteq \cdots \supseteq A$, with $A = \cap_n A_n$. Then, the sequence $\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots$ decreases monotonely to $\mathbb{P}(A)$.

7. *Boole's inequality*: $\mathbb{P}(\cup_i A_i) \leqslant \sum_i \mathbb{P}(A_i)$.

8. *Borel–Cantelli*: Let $A_1, A_2, \ldots$ be a sequence of events, and let $\limsup A_n = \cap_m \cup_{n \geqslant m} A_n$ denote the event that infinitely many $A_n$ occur. Then,

$$\sum_n \mathbb{P}(A_n) < \infty \quad \Rightarrow \quad \mathbb{P}(\limsup A_n) = 0 .$$

   Under the additional assumption that the $\{A_i\}$ are *pairwise independent*,

☞ 616

$$\sum_n \mathbb{P}(A_n) = \infty \quad \Rightarrow \quad \mathbb{P}(\limsup A_n) = 1 .$$

## A.2 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

It is often convenient to describe a random experiment via **random variables**, representing numerical measurements of the experiment. Random variables are usually denoted by capital letters from the last part of the alphabet. A vector $\mathbf{X} = (X_1, \ldots, X_n)$ of random variables is called a **random vector**. A collection of random variables $\{X_t, t \in \mathscr{T}\}$, where $\mathscr{T}$ is any index set, is called a **stochastic process**. The set of possible values for $X_t$ (assuming this is independent of $t$)

is called the **state space** of the process. Stochastic processes are discussed in Sections A.9–A.13. Chapter 5 is devoted to random process generation. ☞ 153

From a mathematical point of view, a random variable $X$ taking values in some set $E$ is a function $X : \Omega \to E$ such that

$$\{X \in B\} \overset{\text{def}}{=} \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{H} \quad \text{for all } B \in \mathcal{E} ,$$

where $\mathcal{E}$ is a $\sigma$-algebra on $E$. The pair $(E, \mathcal{E})$ is called a **measurable space**. If not otherwise specified we assume that $X$ is a **numerical** random variable; that is, $E = \mathbb{R}$. It is sometimes useful to have $E$ as the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. In either case, $\mathcal{E}$ is the corresponding Borel $\sigma$-algebra. The **Borel $\sigma$-algebra** is the smallest $\sigma$-algebra on $\mathbb{R}$ or $\overline{\mathbb{R}}$ that contains all intervals (or, equivalently, all open sets). Elements of this $\sigma$-algebra are called **Borel sets** — for example, a countable union of intervals is a Borel set. The **Lebesgue measure** $m$ is the unique measure on the Borel $\sigma$-algebra such that $m([a, b]) = b - a$. Similar definitions hold for $n$-dimensional Euclidean spaces, replacing intervals by rectangles, etc.

Define

$$P_X(B) = \mathbb{P}(X \in B), \quad B \in \mathcal{E} .$$

Then, $P_X$ is a probability measure on $(E, \mathcal{E})$. It is called the **distribution** of $X$. The probability distribution $P_X$ of a numerical random variable $X$ is completely determined by its **cumulative distribution function** (cdf), defined by

$$F(x) = P_X([-\infty, x]) = \mathbb{P}(X \leqslant x), \quad x \in \overline{\mathbb{R}} .$$

The following properties of a cdf $F$ are a direct consequence of the Kolmogorov axioms. For proofs see, for example, [25].

1. *Right-continuous*: $\lim_{h \downarrow 0} F(x + h) = F(x)$.

2. *Increasing*: $x \leqslant y \Rightarrow F(x) \leqslant F(y)$.

3. *Bounded*: $0 \leqslant F(x) \leqslant 1$.

Conversely, to each function $F$ satisfying the above properties corresponds exactly one distribution $P_X$; see, for example, [5, Theorem 2.2.2]. Figure A.2 shows a generic cdf.

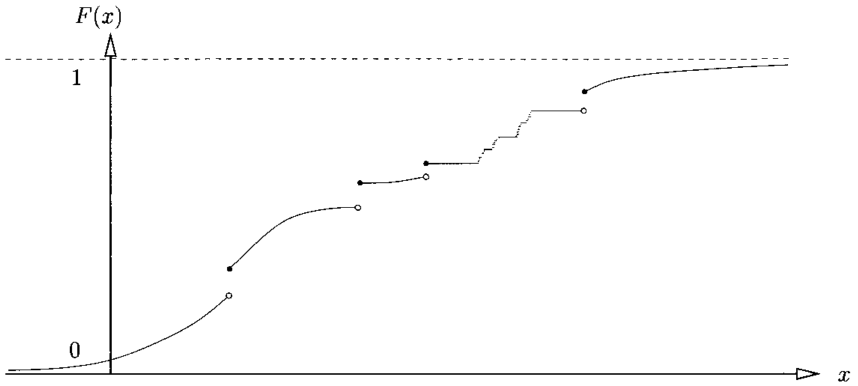

**Figure A.2** A cumulative distribution function (cdf).

A cdf $F_d$ is called **discrete** if there exist numbers $x_1, x_2, \ldots$ and probabilities $0 < f(x_i) \leqslant 1$ summing up to 1, such that for all $x$

$$F_d(x) = \sum_{x_i \leqslant x} f(x_i) . \tag{A.3}$$

Such a cdf is piecewise constant and has jumps of sizes $f(x_1), f(x_2), \ldots$ at points $x_1, x_2, \ldots$, respectively.

A cdf $F_c$ is called **absolutely continuous** if there exists a positive function $f$ such that for all $x$

$$F_c(x) = \int_{-\infty}^{x} f(u)\, \mathrm{d}u . \tag{A.4}$$

Note that such an $F_c$ is differentiable (and hence continuous) with derivative $f$. However, in general the derivative $F_c'$ of a continuous cdf $F_c$ does not necessarily satisfy (A.4). A typical example of a continuous cdf whose derivative is 0 almost everywhere — and hence violates (A.4) — is the **Cantor function**, depicted in Figure A.3. Such continuous cdfs are said to be **singular**. Most distributions used in practice are either discrete or absolutely continuous, or a mixture thereof.

### ■ EXAMPLE A.2 (Cantor Function)

The Cantor function is constructed in the following way. Let $F(1) = 1$. Divide the interval $[0, 1)$ into three equal parts: $[0, \frac{1}{3})$, $[\frac{1}{3}, \frac{2}{3})$, and $[\frac{2}{3}, 1)$. Define $F(x) = \frac{1}{2}$ for $x \in [\frac{1}{3}, \frac{2}{3})$. Next, divide $[0, \frac{1}{3})$ into three subintervals $[0, \frac{1}{9}), [\frac{1}{9}, \frac{2}{9})$, and $[\frac{2}{9}, \frac{3}{9})$ and divide $[\frac{2}{3}, 1)$ into $[\frac{6}{9}, \frac{7}{9}), [\frac{7}{9}, \frac{8}{9})$, and $[\frac{8}{9}, 1)$. Let $F$ have the value $\frac{1}{4}$ on $[\frac{1}{9}, \frac{2}{9})$ and $\frac{3}{4}$ on $[\frac{7}{9}, \frac{8}{9})$. Now divide each of the four remaining subintervals again into three parts. Assign the values $\frac{1}{8}, \frac{3}{8}, \frac{5}{8}$, and $\frac{7}{8}$ to the middle intervals, and continue this process indefinitely. This cdf is continuous, but its derivative is 0 almost everywhere.

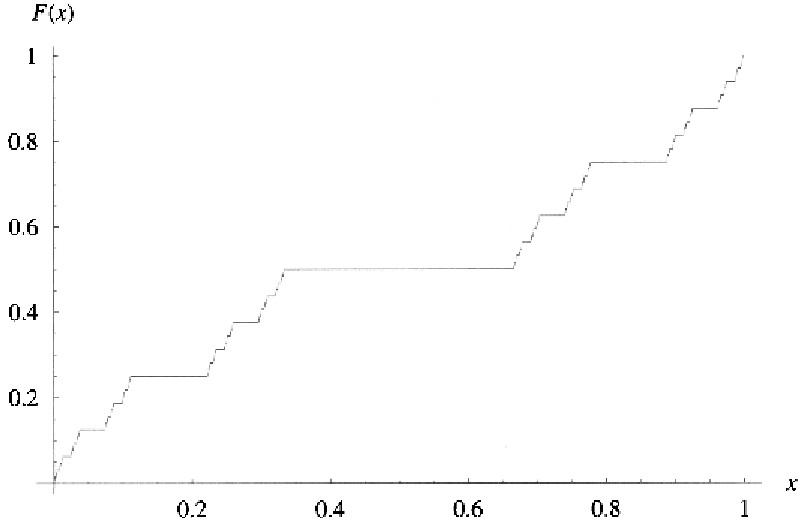

**Figure A.3** The Cantor function is a continuous singular cdf.

It can be shown (see, for example, [5, Chapter 1]) that every cdf $F$ can be written as the unique convex combination, or **mixture**, of a discrete, an absolutely continuous, and a continuous singular cdf:

$$F(x) = \alpha_1 F_d(x) + \alpha_2 F_c(x) + \alpha_3 F_s(x), \quad \text{where} \quad \alpha_1 + \alpha_2 + \alpha_3 = 1\,,$$

and $\alpha_k \geqslant 0$ for $k = 1, 2, 3$.

### A.2.1 Probability Density

A probability distribution on some measurable space $(E, \mathcal{E})$ is often of the form

$$P_X(B) = \int_B f(x)\,\mathrm{d}m(x), \quad B \in \mathcal{E}\,,$$

where $m$ is some measure on $(E, \mathcal{E})$. We say that $P_X$ has a **probability density function** (pdf), or simply **density**, $f$ with respect to $m$.

■ **EXAMPLE A.3 (Discrete Distribution)**

Suppose a random variable $X$ has a discrete cdf, as in (A.3). Thus, $X$ takes values in some finite or countable set of points $E = \{x_1, x_2, \ldots\}$, with $\mathbb{P}(X = x_i) = f(x_i) > 0$, $i = 1, 2, \ldots$. Define $f(x) = 0$ for all $x \notin E$. Let $\mathcal{E}$ denote the collection of all subsets of $E$ and let $m$ be the **counting measure** on $(E, \mathcal{E})$, that is, $m(B)$ is the number of points in $E$ that lie in the set $B$. Then, we see that the distribution $P_X$ of $X$ satisfies

$$P_X(B) = \mathbb{P}(X \in B) = \sum_{x_i \in B} f(x_i) = \int_B f(x)\,\mathrm{d}m(x) \;\text{ for all } B \subseteq E\,. \tag{A.5}$$

In other words, $X$ has a density $f$ with respect to the counting measure $m$. Such a random variable is called **discrete** and $P_X$ is called a **discrete distribution**. Such a distribution is thus completely specified by its (discrete) pdf, and probabilities can be evaluated via summation, as in (A.5). This is illustrated in Figure A.4. Many specific discrete distributions are given in Chapter 4. ☞ 85

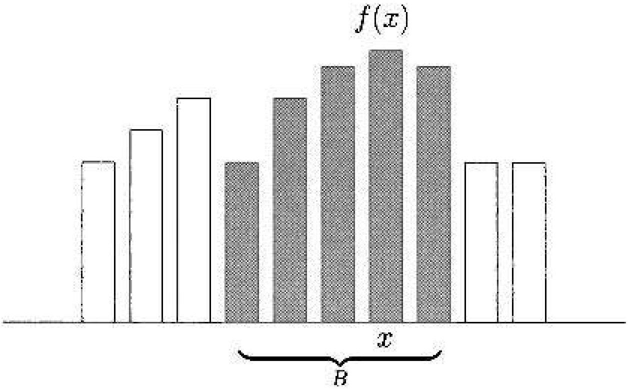

**Figure A.4** Discrete probability density function (pdf). The shaded area corresponds to the probability $\mathbb{P}(X \in B)$.

### ■ EXAMPLE A.4 (Absolutely Continuous Distribution)

Suppose a random variable $X$ has an absolutely continuous cdf, as in (A.4). Then, the distribution $P_X$ of $X$ satisfies

$$P_X(B) = \mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x = \int_B f(x)\,\mathrm{d}m(x) \tag{A.6}$$

for all Borel sets $B$, where $m$ is the Lebesgue measure. The distribution $P_X$ is said to be **absolutely continuous** with respect to the Lebesgue measure, and $f$ is the corresponding pdf. As a consequence, such a distribution is completely specified by its pdf, and probabilities can be evaluated via integration. This is illustrated in Figure A.5. Many specific absolutely continuous distributions are given in Chapter 4. ☞ 85

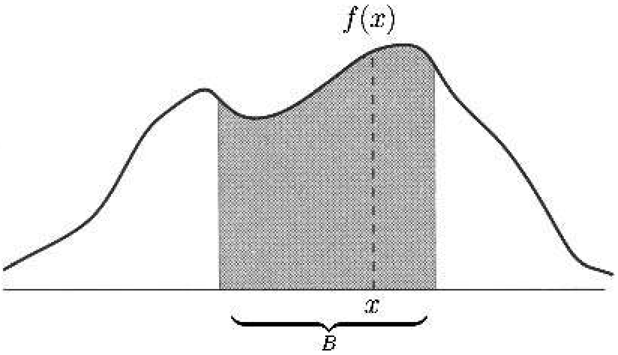

**Figure A.5** Absolutely continuous probability density function (pdf). The shaded area corresponds to the probability $\mathbb{P}(X \in B)$.

We can view $f(x)$ as the probability "density" at $X = x$, in the sense that, for small $h$,

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_x^{x+h} f(u)\,\mathrm{d}u \approx h\,f(x)\,.$$

**Remark A.2.1 (Probability Density and Probability Mass)** It is important to note that we deliberately use the *same* name, "pdf", and symbol, $f$, in both the discrete and the absolutely continuous case, rather than distinguish between a probability mass function (pmf) and probability density function (pdf). The reason is that from a measure-theoretic point of view the pdf plays exactly the same role in the discrete and absolutely continuous cases. The only difference is the measure $m$. We use the notation $X \sim \mathsf{Dist}$, $X \sim f$, and $X \sim F$ to indicate that $X$ has distribution $\mathsf{Dist}$, pdf $f$, and cdf $F$.

## A.2.2 Joint Distributions

Distributions for random vectors and stochastic processes can be specified in much the same way as for random variables. In particular, the distribution of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is completely determined by specifying the **joint cdf** $F$, defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n), \quad x_i \in \mathbb{R},\ i = 1, \ldots, n\,.$$

Similarly, the distribution of a stochastic process $\{X_t, t \in \mathscr{T}\}$, with $\mathscr{T} \subseteq \mathbb{R}$, is completely determined by its **finite-dimensional distributions**; that is, the distributions of the random vectors $(X_{t_1}, \ldots, X_{t_n})$ for any choice of $n$ and $t_1, \ldots, t_n$.

By analogy to the one-dimensional case, a random vector $\mathbf{X}$ taking values in $\mathbb{R}^n$ is said to have a pdf $f$ with respect to some measure $m$, if

$$\mathbb{P}(\mathbf{X} \in B) = \int_B f(\mathbf{x})\, \mathrm{d}m(\mathbf{x})\,, \tag{A.7}$$

for all $n$-dimensional Borel sets $B$. The important cases are when $m$ is either the counting measure or the Lebesgue measure.

The **marginal pdfs** can be recovered from the joint pdf by "integrating out the other variables". For example, for a random vector $(X, Y)$ with pdf $f$ with respect to the Lebesgue measure on $\mathbb{R}^2$, the pdf $f_X$ of $X$ is given by

$$f_X(x) = \int f(x, y)\, \mathrm{d}y\,.$$

**Remark A.2.2 (Multivariate Singular Distributions)** Continuous singular distributions are much more likely to be encountered in the multidimensional setting. For example, if a numerical random vector takes values exclusively in a lower dimensional subset, then the distribution has a derivative of 0 almost everywhere with respect to the Lebesgue measure and so is singular.

## A.3 EXPECTATION AND VARIANCE

It is often useful to consider various numerical characteristics of a random variable or its distribution. For example, two such quantities are the expectation and variance. The first measures the mean value of the distribution; the second measures the spread or dispersion of the distribution. The intuitive definition of the expectation of a discrete random variable $X$ is that it is the average of the possible values that $X$ can take, weighted by the corresponding probabilities; that is,

$$\mathbb{E}X = \sum_x x\, \mathbb{P}(X = x)\,.$$

Similarly, for an absolutely continuous random variable $X$ the expectation is given by

$$\mathbb{E}X = \int x\, f(x)\, \mathrm{d}x\,.$$

Both definitions are part of a more general framework, in which the **expectation** of $X$ is defined as the abstract integral

$$\mathbb{E}X = \int X\, \mathrm{d}\mathbb{P}\,, \tag{A.8}$$

which is defined in four steps (see, for example, [5, Chapter 3] and [12]):

1. If $X$ is an **indicator function** of some event $A$, that is,

$$X = \mathrm{I}_A \overset{\text{def}}{=} \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \,, \end{cases} \tag{A.9}$$

then

$$\mathbb{E}X \overset{\text{def}}{=} \mathbb{P}(A) \,. \tag{A.10}$$

2. If $X$ is positive and **simple**, that is, $X = \sum_{i=1}^{n} a_i \, \mathrm{I}_{A_i}$ for some positive (possibly infinite) numbers $a_1, \ldots, a_n$ and events $A_1, \ldots, A_n$, then

$$\mathbb{E}X \overset{\text{def}}{=} \sum_{i=1}^{n} a_i \, \mathbb{P}(A_i) \,, \tag{A.11}$$

with $\infty \times 0$ and $0 \times \infty$ defined to be 0.

3. If $X$ is a positive random variable, then

$$\mathbb{E}X \overset{\text{def}}{=} \lim_{n \to \infty} \mathbb{E}X_n \,, \tag{A.12}$$

where $X_1, X_2, \ldots$ is any sequence of simple random variables that increases almost surely to $X$. We write ☞ 623

$$X_n \overset{a.s.}{\nearrow} X \,.$$

It can be shown [12] that such a sequence exists, and that the limit in (A.12) exists (possibly infinity) and does not depend on the increasing sequence of simple random variables.

4. Finally, for a general (not necessarily positive) random variable $X$, write $X = X^+ - X^-$, where $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$ are the positive and negative parts of $X$ (note that both are $\geqslant 0$), and define

$$\mathbb{E}X \overset{\text{def}}{=} \mathbb{E}X^+ - \mathbb{E}X^- \,,$$

*provided* that at least one of the right-hand-side terms is finite ($\infty - \infty$ is not well-defined).

Random variables $X$ for which $\mathbb{E}|X| < \infty$ (and hence the expectation is finite) are called **integrable**. It is not difficult to show [3, Page 216] that a random variable $X$ is integrable if and only if

$$\lim_{c \to \infty} \mathbb{E}|X| \, \mathrm{I}_{\{|X| > c\}} = 0 \,.$$

A random variable $X$ is said to be **square integrable** if $\mathbb{E}X^2 < \infty$. A sequence of random variables $X_1, X_2, \ldots$ is said to be **integrable** if

$$\sup_n \mathbb{E}|X_n| < \infty \,.$$

A sequence of random variables $X_1, X_2, \ldots$ is said to be **uniformly integrable**, if

$$\lim_{c \to \infty} \sup_n \mathbb{E}|X_n| \, \mathrm{I}_{\{|X_n| > c\}} = 0 \,.$$

In particular, $X_1, X_2, \ldots$ must be integrable. Moreover, if for some $\varepsilon > 0$

$$\sup_n \mathbb{E}|X_n|^{1+\varepsilon} < \infty \, ,$$

then the sequence $X_1, X_2, \ldots$ is uniformly integrable. Another sufficient condition for uniform integrability [2, Page 32] is the existence of an integrable random variable $Y$ such that $\mathbb{P}(|X_n| \geqslant x) \leqslant \mathbb{P}(|Y| \geqslant x)$ for all $x$ and $n$. For continuous-time stochastic processes $\{X_t, t \geqslant 0\}$, integrability and uniform integrability are defined in the same way, replacing the discrete $n$ with a continuous $t$.

For the purpose of calculating expectations, the following theorem is indispensable.

**Theorem A.3.1 (Expected Value)** Let $X$ be a random variable with distribution $P_X$ and cdf $F$, and let $g$ be a numerical function, then (provided that the integral exists)

$$\mathbb{E}\, g(X) = \int g(X)\, \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g(x)\, \mathrm{d}P_X(x) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} g(x)\, \mathrm{d}F(x) \, . \qquad \text{(A.13)}$$

The last integral in (A.13) is called a **Lebesgue–Stieltjes** integral. In most cases of practical interest this integral can be determined via elementary summation or Riemann integration, in which it can be viewed as a Riemann–Stieltjes integral [3, Page 228]. In particular, when $X$ is discrete with pdf $f$, (A.13) reduces to

$$\mathbb{E}\, g(X) = \sum_x g(x) f(x) \, , \qquad \text{(A.14)}$$

and in the absolutely continuous case (A.13) becomes

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x)\, f(x)\, \mathrm{d}x \, . \qquad \text{(A.15)}$$

Theorem A.3.1 can be readily generalized to random vectors. In particular, if $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random vector with ($n$-dimensional) cdf $F$, and $g$ a numerical function on $\mathbb{R}^n$, then

$$\mathbb{E}\, g(\mathbf{X}) = \int_{\mathbb{R}^n} g(\mathbf{x})\, \mathrm{d}F(\mathbf{x}) \, . \qquad \text{(A.16)}$$

### A.3.1 Properties of the Expectation

☞ 623

Below, $X$, $X_1, X_2, \ldots$, and $Y$ are random variables, and $\mathbf{X}$ is a random vector. We write $X_n \xrightarrow{\text{a.s.}} X$ to indicate that the sequence $X_1, X_2, \ldots$ converges almost surely to $X$. Note that Properties 1, 2, 4, and 7 below follow directly from the definition of the expectation. Proofs of the other properties can be found, for example, in [3, Chapter 3]. See also Section A.8.

1. *Positivity*: For positive random variables the expectation always exists (possibly $+\infty$).

2. *Linearity*: $\mathbb{E}(aX + bY) = a\,\mathbb{E}X + b\,\mathbb{E}Y$ for $a, b \in \mathbb{R}$.

3. *Monotonicity*: If $X \geqslant Y$, then $\mathbb{E}X \geqslant \mathbb{E}Y$.

4. *Indicator*: If $\mathrm{I}_A$ is the indicator of the event $A$, then $\mathbb{E}\,\mathrm{I}_A = \mathbb{P}(A)$.

5. *Jensen's inequality*: Let $\mathscr{X}$ be a convex subset of $\mathbb{R}^n$ and $h : \mathscr{X} \to \mathbb{R}$ be a convex measurable function. Let $\mathbf{X}$ be a random vector taking values in $\mathscr{X}$, such that $\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)$ is finite. Then, $\mathbb{E}h(\mathbf{X})$ exists and ☞ 679

$$\mathbb{E}h(\mathbf{X}) \geqslant h(\mathbb{E}\mathbf{X}) .$$

6. *Fatou's lemma*: If $X_n \geqslant 0$, then

$$\mathbb{E} \liminf_{n\to\infty} X_n \leqslant \liminf_{n\to\infty} \mathbb{E}X_n .$$

7. *Monotone convergence theorem*: Suppose $\mathbb{E}X_n$ exists for some $n$, then

$$X_n \overset{a.s.}{\nearrow} X \quad \Rightarrow \quad \mathbb{E}X_n \nearrow \mathbb{E}X .$$

8. *Dominated convergence theorem*: Suppose $|X_n| \leqslant Y$ for all $n$, where $\mathbb{E}Y < \infty$. Then,

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad \mathbb{E}X_n \to \mathbb{E}X .$$

### A.3.2 Variance

The **variance** of a random variable $X$, denoted by $\mathrm{Var}(X)$ (or sometimes $\sigma^2$) is defined by

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 .$$

The square root of the variance is called the **standard deviation.**

In general, the mean and the variance do not give enough information to completely specify the distribution of a random variable. However, they may provide useful bounds. We give three such bounds. A proof of Kolmogorov's inequality may be found in [5, Page 116].

1. *Markov's inequality*: For any positive random variable $X$ with expectation $\mu$,

$$\mathbb{P}(X \geqslant x) \leqslant \frac{\mu}{x}, \quad x \geqslant 0 . \tag{A.17}$$

2. *Chebyshev's inequality*: Let $X$ be a random variable with finite expectation and variance, $\mu$ and $\sigma^2$, respectively. Then,

$$\mathbb{P}(|X - \mu| \geqslant x) \leqslant \frac{\sigma^2}{x^2}, \quad x \geqslant 0 . \tag{A.18}$$

3. *Kolmogorov's inequality*: Let $X_1, X_2, \ldots$ be a sequence of independent random variables. Let $S_1, S_2, \ldots$ be the sequence of partial sums, defined by $S_n = X_1 + \cdots + X_n$ and assumed to have finite expectations and variances, $\{\mu_n\}$ and $\{\sigma_n^2\}$, respectively. Then,

$$\mathbb{P}\left(\max_{1\leqslant i\leqslant n} |S_i - \mu_i| \geqslant x\right) \leqslant \frac{\sigma_n^2}{x^2}, \quad x \geqslant 0 . \tag{A.19}$$

## A.4 CONDITIONING AND INDEPENDENCE

Conditional probabilities and conditional distributions are used to model additional information on a random experiment. Independence is used to model *lack* of such information.

### A.4.1 Conditional Probability

Suppose some event $B \subseteq \Omega$ occurs. Given this fact, event $A$ will occur if and only if $A \cap B$ occurs, and the relative chance of $A$ occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$, provided $\mathbb{P}(B) > 0$. This leads to the definition of the **conditional probability** of $A$ given $B$:

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{if } \mathbb{P}(B) > 0 . \tag{A.20}$$

The above definition breaks down if $\mathbb{P}(B) = 0$. Such conditional probabilities must be treated with more care [3].

Three important consequences of the definition of conditional probability are:

1. *Product rule*: For any sequence of events $A_1, A_2, \ldots, A_n$,

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}) , \tag{A.21}$$

   using the abbreviation $A_1 A_2 \cdots A_k = A_1 \cap A_2 \cap \cdots \cap A_k$.

2. *Law of total probability*: If $\{B_i\}$ forms a **partition** of $\Omega$ (that is, $B_i \cap B_j = \emptyset, i \neq j$ and $\cup_i B_i = \Omega$), then for any event $A$

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i) . \tag{A.22}$$

3. *Bayes' rule*: Let $\{B_i\}$ form a partition of $\Omega$. Then, for any event $A$ with $\mathbb{P}(A) > 0$,

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\,\mathbb{P}(B_j)}{\sum_i \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i)} . \tag{A.23}$$

### A.4.2 Independence

Two events $A$ and $B$ are said to be **independent** if the knowledge that $B$ has occurred does not change the probability that $A$ occurs. That is, $A$, $B$ independent $\Leftrightarrow \mathbb{P}(A \,|\, B) = \mathbb{P}(A)$. Since $\mathbb{P}(A \,|\, B)\,\mathbb{P}(B) = \mathbb{P}(A \cap B)$, an alternative definition of independence is

$$A,\ B \text{ independent} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) .$$

This definition covers the case where $\mathbb{P}(B) = 0$ and can be extended to arbitrarily many events: events $A_1, A_2, \ldots$ are said to be **independent** if for any $k$ and any choice of distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) .$$

The $\{A_i\}$ are said to be **pairwise independent** if every two events are independent.

The concept of independence can also be formulated for *random variables.* Random variables $X_1, X_2, \ldots$ are said to be **independent** if the events $\{X_{i_1} \in B_1\}, \ldots, \{X_{i_n} \in B_n\}$ are independent for all finite choices of $n$, distinct indices $i_1, \ldots, i_n$, and Borel sets $B_1, \ldots, B_n$.

An important characterization of independent random variables is the following (for a proof, see [25], for example).

**Theorem A.4.1 (Product of Marginal Pdfs)** *Random variables $X_1, \ldots, X_n$ with marginal pdfs $f_{X_1}, \ldots, f_{X_n}$ and joint pdf $f$ are independent if and only if*

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \quad \textit{for all } x_1, \ldots, x_n \, . \tag{A.24}$$

Many probabilistic models involve random variables $X_1, X_2, \ldots$ that are **independent and identically distributed**, abbreviated as **iid**. We will use this abbreviation throughout this book.

### A.4.3 Covariance

The **covariance** of two random variables $X$ and $Y$ with expectations $\mu_X$ and $\mu_Y$, respectively, is defined as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \, .$$

This is a measure for the amount of linear dependency between the variables. Let $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$. A scaled version of the covariance is given by the **correlation coefficient**,

$$\varrho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \, \sigma_Y} \, .$$

Below we use the notation $\mu_X = \mathbb{E}X$ and $\sigma_X^2 = \mathrm{Var}(X)$. The following properties follow directly from the definitions of variance and covariance.

1. $\mathrm{Var}(X) = \mathbb{E}X^2 - \mu_X^2$.

2. $\mathrm{Var}(aX + b) = a^2 \, \sigma_X^2$.

3. $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X \, \mu_Y$.

4. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

5. $-\sigma_X \sigma_Y \leqslant \mathrm{Cov}(X,Y) \leqslant \sigma_X \sigma_Y$.

6. $\mathrm{Cov}(aX + bY, Z) = a \, \mathrm{Cov}(X,Z) + b \, \mathrm{Cov}(Y,Z)$.

7. $\mathrm{Cov}(X,X) = \sigma_X^2$.

8. $\mathrm{Var}(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2\,\mathrm{Cov}(X,Y)$.

9. If $X$ and $Y$ are independent, then $\mathrm{Cov}(X,Y) = 0$.

As a consequence of Properties 2 and 8 we have that for any sequence of *independent* random variables $X_1$, ..., $X_n$ with variances $\sigma_1^2, \ldots, \sigma_n^2$,

$$\mathrm{Var}(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2 \ , \tag{A.25}$$

for any choice of constants $a_1, \ldots, a_n$.

For random vectors, such as $\mathbf{X} = (X_1, \ldots, X_n)$, it is convenient to write the expectations and covariances in vector notation. It will usually be clear from the context whether we interpret $\mathbf{X}$ as a *row* or a *column vector*. In some cases, for example, with matrix multiplication, we make the distinction explicit. For a random (column) vector $\mathbf{X}$ we define its **expectation vector** as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^\top \ .$$

The **covariance matrix** $\Sigma$ is defined as the matrix whose $(i,j)$-th element is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \ .$$

If we define the expectation of a vector (matrix) to be the vector (matrix) of expectations, then we can compactly write

$$\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$$

and

$$\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] \ .$$

### A.4.4 Conditional Density and Expectation

Suppose $X$ and $Y$ are both discrete or both absolutely continuous, with joint pdf $f$, and suppose $f_X(x) > 0$. Then, the **conditional pdf** of $Y$ given $X = x$ is given by

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} \quad \text{for all } y \ . \tag{A.26}$$

In the discrete case the formula is a direct translation of (A.20), with $f_{Y|X}(y \mid x) = \mathbb{P}(Y = y \mid X = x)$. In the absolutely continuous case a similar interpretation, in terms of densities, can be used (see, for example, [25, Page 221]). The corresponding distribution is called the **conditional distribution** of $Y$ given $X = x$, and the corresponding **conditional expectation** is

$$\mathbb{E}[Y \mid X = x] = \begin{cases} \sum_y y \, f_{Y|X}(y \mid x) & \text{discrete case,} \\ \int y \, f_{Y|X}(y \mid x) \, \mathrm{d}y & \text{absolutely continuous case.} \end{cases} \tag{A.27}$$

Note that $\mathbb{E}[Y \mid X = x]$ is a function of $x$. The corresponding random variable is written as $\mathbb{E}[Y \mid X]$. A similar formalism can be used when conditioning on a sequence of random variables $X_1, \ldots, X_n$ or on a $\sigma$-algebra; see, for example, [5, Chapter 9]. The conditional expectation has similar properties to the ordinary expectation in Section A.3.1. Other useful properties (see, for example, [28]) are:

1. *Tower property*: If $\mathbb{E}Y$ exists, then

$$\mathbb{E}\,\mathbb{E}[Y \mid X] = \mathbb{E}Y \ . \tag{A.28}$$

2. *Taking out what is known*: If $\mathbb{E}Y$ exists, then
$$\mathbb{E}[XY \mid X] = X\mathbb{E}Y .$$

3. *Orthogonal projection*: If $Y$ is square integrable, then $\mathbb{E}[Y|X]$ is the function $h(X)$ that minimizes $\mathbb{E}(Y - h(X))^2$.

## A.5 $L^P$ SPACE

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $X$ a numerical random variable. For $p \in [1, \infty)$ define
$$\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}$$
and let
$$\|X\|_\infty = \inf\{x : \mathbb{P}(|X| \leqslant x) = 1\} .$$
For each $p \in [1, \infty]$ we denote by $L^p$ the collection of all numerical random variables $X$ for which $\|X\|_p < \infty$. In particular $L^1$ is comprised of all integrable random variables and $L^2$ is comprised of all square integrable random variables.

The following properties of $L^p$ spaces can be found, for example, in [26, Chapter 3].

1. *Positivity*: $\|X\|_p \geqslant 0$, and $\|X\|_p = 0 \Leftrightarrow X = 0$ (a.s.).
2. *Multiplication with a constant*: $\|c\,X\|_p = |c|\,\|X\|_p$.
3. *Minkowski's (triangle) inequality*: $\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p$.
4. *Hölder's inequality*: For $p, q, r \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$,
$$\|XY\|_r \leqslant \|X\|_p\,\|Y\|_q . \tag{A.29}$$
The particular case with $p = q = 2$ and $r = 1$ is called **Schwarz's inequality**.

5. *Monotonicity*: If $1 \leqslant p < q \leqslant \infty$, then $\|X\|_p \leqslant \|X\|_q$.

The space $L^p$ is a *linear space*. The first three properties above identify $\|\cdot\|_p$ as a norm on this space, provided that random variables that are almost surely equal are identified as one and the same. Of particular importance is $L^2$, which is in fact a *Hilbert space*, with inner product
$$\langle X, Y\rangle = \mathbb{E}[XY] .$$
We denote the $L^2$ norm simply by $\|\cdot\|$, suppressing the subscript.

For random variables in $L^2$ the concepts of variance and covariance have a geometric interpretation. Namely, if $X$ and $Y$ are zero-mean random variables (their expectation is 0), then
$$\mathrm{Var}(X) = \|X\|^2 \quad \text{and} \quad \mathrm{Cov}(X, Y) = \langle X, Y\rangle .$$
Another important use of $L^2$ spaces is in conditioning. Let $X$ and $Y$ be random variables. Define $\mathcal{K}$ to be the space of functions of $X$ that are square integrable.

There exists a unique (up to equivalence) element in $\mathcal{K}$ that solves the minimization program

$$\min_{K \in \mathcal{K}} \|Y - K\| .$$

This is the **orthogonal projection** of $Y$ onto $\mathcal{K}$, and it coincides (up to equivalence) with the conditional expectation $\mathbb{E}[Y \mid X]$; see, for example, [28, Chapter 9].

## A.6 FUNCTIONS OF RANDOM VARIABLES

### A.6.1 Linear Transformations

Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be a column vector in $\mathbb{R}^n$ and $A$ an $m \times n$ matrix. The mapping $\mathbf{x} \mapsto \mathbf{z}$, with $\mathbf{z} = A\mathbf{x}$, is called a **linear transformation**. Now consider a random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and let

$$\mathbf{Z} = A\mathbf{X} .$$

Then $\mathbf{Z}$ is a random vector in $\mathbb{R}^m$. If $\mathbf{X}$ has an expectation vector $\boldsymbol{\mu}_\mathbf{X}$ and covariance matrix $\Sigma_\mathbf{X}$, then the expectation vector of $\mathbf{Z}$ is

$$\boldsymbol{\mu}_\mathbf{Z} = A\boldsymbol{\mu}_\mathbf{X} \tag{A.30}$$

and the covariance matrix of $\mathbf{Z}$ is

$$\Sigma_\mathbf{Z} = A\, \Sigma_\mathbf{X}\, A^\top . \tag{A.31}$$

If, moreover, $A$ is an invertible $n \times n$ matrix and $\mathbf{X}$ has a pdf $f_\mathbf{X}$, then the pdf of $\mathbf{Z}$ is given by

$$f_\mathbf{Z}(\mathbf{z}) = \frac{f_\mathbf{X}(A^{-1}\mathbf{z})}{|\det(A)|}, \quad \mathbf{z} \in \mathbb{R}^n , \tag{A.32}$$

where $|\det(A)|$ denotes the absolute value of the determinant of $A$.

### A.6.2 General Transformations

For a generalization of the linear transformation rule (A.32), consider an arbitrary mapping $\mathbf{x} \mapsto \boldsymbol{g}(\mathbf{x})$, written out:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix} .$$

For a fixed $\mathbf{x}$, let $\mathbf{z} = \boldsymbol{g}(\mathbf{x})$. Suppose that the inverse mapping $\boldsymbol{g}^{-1}$ of $\boldsymbol{g}$ exists; hence, $\mathbf{x} = \boldsymbol{g}^{-1}(\mathbf{z})$. Let $\mathbf{X}$ be a random vector with pdf $f_\mathbf{X}$, and let $\mathbf{Z} = \boldsymbol{g}(\mathbf{X})$. Then, $\mathbf{Z}$ has pdf

$$f_\mathbf{Z}(\mathbf{z}) = \frac{f_\mathbf{X}(\mathbf{x})}{|\det(J_{\boldsymbol{g}}(\mathbf{x}))|}, \quad \mathbf{z} \in \mathbb{R}^n , \tag{A.33}$$

☞ 710 where $J_{\boldsymbol{g}}(\mathbf{x})$ is the Jacobi matrix at $\mathbf{x}$ of the transformation $\boldsymbol{g}$.

**Remark A.6.1 (Coordinate Transformation)** Typically, in coordinate transformations it is $\boldsymbol{g}^{-1}$ that is given — that is, an expression for $\mathbf{x}$ as a function of $\mathbf{z}$, rather than $\boldsymbol{g}$. Note that $|\det(J_{\boldsymbol{g}^{-1}}(\mathbf{z}))| = 1/|\det(J_{\boldsymbol{g}}(\mathbf{x}))|$.

## A.7 GENERATING FUNCTION AND INTEGRAL TRANSFORMS

Many calculations and manipulations involving probability distributions are facilitated by the use of transform techniques. All such transforms share two important properties:

1. *Uniqueness*: Two distributions are the same if and only if their respective transforms are the same.
2. *Independence*: If $X$ and $Y$ are independent with transform $T_X$ and $T_Y$, respectively, then the transform $T_{X+Y}$ of $X+Y$ is given by the product

$$T_{X+Y}(t) = T_X(t)\,T_Y(t)\ .$$

In this section the $k$-th derivative of a function $g$ is denoted by $g^{(k)}$.

### A.7.1 Probability Generating Function

Let $X$ be a random variable taking values in some subset of the positive integers, $\mathbb{N} = \{0, 1, 2, \dots\}$, with discrete pdf $f$. The **probability generating function** of $X$ is the function $G$ defined by

$$G(z) = \mathbb{E}\, z^X = \sum_{x=0}^{\infty} z^x\, f(x)\,.$$

The power series that defines $G$ converges for all $|z| \leqslant r$, for some $r \geqslant 1$. Two useful properties are (see, for example, [9, Chapter XI]):

1. *Inversion*: $f(x) = \dfrac{G^{(x)}(0)}{x!}$, $x \in \mathbb{N}$.
2. *Moment property*: $\mathbb{E}[X(X-1)\cdots(X-k+1)] = \lim_{z\uparrow 1} G^{(k)}(z)$, $k = 1, 2, \dots$.

### A.7.2 Moment Generating Function and Laplace Transform

The **moment generating function** of a random variable $X$ with cdf $F$ is the function $M : \mathbb{R} \to [0, \infty]$, given by

$$M(t) = \mathbb{E}\,\mathrm{e}^{tX} = \int_{-\infty}^{\infty} \mathrm{e}^{tx}\,\mathrm{d}F(x)\ .$$

Note that the expectation always exists, but can be $+\infty$. For a *positive* random variable $X$ its **Laplace transform** is the function $L : \mathbb{R}_+ \to \mathbb{R}_+$, defined by $L(t) = M(-t)$, $t \geqslant 0$. When $X$ has an absolutely continuous distribution with pdf $f$, the Laplace transform coincides with the classical Laplace transform of the function $f$.

If the moment generating function is finite in an open interval containing 0, then the integer moments $\{\mathbb{E}X^k\}$ exist, are finite, and uniquely determine the distribution of $X$. Moreover, in that case the following properties hold (see, for example, [5]):

1. *Moment property*: $\mathbb{E}X^k = M^{(k)}(0),\ k \geqslant 1$.
2. *Taylor's theorem*:
$$M(t) = \sum_{k=0}^{\infty} \frac{\mathbb{E}X^k}{k!}\, t^k \,.$$

**Remark A.7.1 (Infinite Moment Generating Function)** If the moment generating function is not finite in any open interval containing 0, then the sequence of integer moments, even if they are all finite, is not sufficient to uniquely characterize the distribution of a random variable; see, for example, [13].

### A.7.3 Characteristic Function

The most general transform concept is that of the characteristic function. Every random variable has a characteristic function. It is closely related to the classical Fourier transform of a function and has superior analytical properties to the moment generating function.

The **characteristic function** of a random variable $X$ with cdf $F$, is the function $\phi : \mathbb{R} \to \mathbb{C}$, defined by

$$\phi(t) = \mathbb{E}\,\mathrm{e}^{\mathrm{i}tX} = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}F(x), \qquad t \in \mathbb{R}\,,$$

or, equivalently,

$$\phi(t) = \mathbb{E}\cos(tX) + \mathrm{i}\,\mathbb{E}\sin(tX), \qquad t \in \mathbb{R}\,.$$

Note that $\phi(0) = 1$ and $|\phi(t)| \leqslant 1$. Some other properties are (for proofs see [5], for example):

1. *Moment property*: If $\mathbb{E}|X|^n < \infty$, then, for $k = 1, 2, \ldots, n$, $\phi^{(k)}$ is finite and continuous on $\mathbb{R}$, with
$$\phi^{(k)}(t) = \mathrm{i}^k\, \mathbb{E}\left[X^k\, \mathrm{e}^{\mathrm{i}tX}\right], \qquad t \in \mathbb{R}\,,$$
and so, in particular, $\mathbb{E}X^k = (-\mathrm{i})^k \phi^{(k)}(0)$.
2. *Taylor's theorem*: If $\mathbb{E}|X|^n < \infty$, then, in a neighborhood of 0,
$$\phi(t) = \sum_{k=0}^{n} \frac{\mathbb{E}X^k}{k!} (\mathrm{i}t)^k + o(t^n)\,.$$
3. *Continuity*: Let $F_1, F_2, \ldots$ be a sequence of cdfs, with characteristic functions $\phi_1, \phi_2, \ldots$. If $\phi_n(t) \to \phi(t)$, pointwise, and $\phi(t)$ is continuous at $t = 0$, then there exists a cdf $F$ such that $F_n$ *converges weakly* (see Page 623) to $F$, and $\phi$ is its characteristic function.

4. $\phi(t)$ is uniformly continuous on $\mathbb{R}$.

5. $\phi_{(-X)}(t) = \overline{\phi_X(t)}$, $t \in \mathbb{R}$, from which it follows that a random variable is *symmetric* around 0 (that is, $X$ and $-X$ are identically distributed) if and only if its characteristic function is *real-valued.*

## A.8 LIMIT THEOREMS

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, and let $X_1, X_2, \ldots, X$ be random variables taking values in a metric space $E$ with distance $\varrho$ and equipped with a $\sigma$-algebra $\mathcal{E}$. A typical example is $E = \mathbb{R}^n$ with $\varrho$ the Euclidean distance $\varrho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$; see also [2]. Recall that for numerical random variables $E = \mathbb{R}$ and $\varrho(x, y) = |x - y|$.

### A.8.1 Modes of Convergence

We have the following definitions of the different modes of convergence of random variables.

1. *Almost sure convergence*: The sequence of numerical random variables $X_1, X_2, \ldots$ is said to converge **almost surely** to a numerical random variable $X$, denoted $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1 \, .$$

2. *Convergence in $L^p$-norm*: The sequence of numerical random variables $X_1, X_2, \ldots$ is said to converge **in $L^p$-norm** to a numerical random variable $X$, denoted $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0 \, ,$$

or, equivalently, if $\lim_{n\to\infty} \|X_n - X\|_p = 0$, where $\|\cdot\|_p$ denotes the $L^p$ norm. Convergence in $L^2$-norm is often called **mean square convergence.** ☞ 619

3. *Convergence in probability*: The sequence $X_1, X_2, \ldots$ is said to converge **in probability** to $X$, denoted $X_n \xrightarrow{\mathbb{P}} X$, if

$$\lim_{n\to\infty} \mathbb{P}(\varrho(X_n, X) < \varepsilon) = 1 \quad \text{for all } \varepsilon > 0 \, .$$

4. *Convergence in distribution*: Let $P_{X_n}$ be the distribution of $X_n$ and $P_X$ the distribution of $X$. The sequence $X_1, X_2, \ldots$ is said to converge **in distribution** to $X$, denoted $X_n \xrightarrow{d} X$, if the distribution $P_{X_n}$ **converges weakly** to $P_X$, that is,

$$\lim_{n\to\infty} P_{X_n}(A) = P_X(A)$$

for all sets $A \in \mathcal{E}$ such that $P_X(\partial A) = 0$, where $\partial A \in \mathcal{E}$ is the boundary of $A$. An equivalent definition is that

$$\lim_{n\to\infty} \mathbb{E}h(X_n) = \mathbb{E}h(X)$$

for all bounded continuous functions $h : E \to \mathbb{R}$.

5. *Complete convergence*: A sequence of random variables $X_1, X_2, \dots$ is said to converge **completely** to $X$, denoted $X_n \xrightarrow{\text{cpl.}} X$, if

$$\sum_n \mathbb{P}(\varrho(X_n, X) > \varepsilon) < \infty \quad \text{for all } \varepsilon > 0 \, .$$

The most general relationships among the various modes of convergence for numerical random variables are shown on the following diagram. Proofs can be found in [2] and [3]. See also [14].

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{cpl.}} X} \arrow[r, Rightarrow] & \boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[d, Rightarrow] & \\
& \boxed{X_n \xrightarrow{\mathbb{P}} X} \arrow[r, Rightarrow] & \boxed{X_n \xrightarrow{d} X} \\
\boxed{X_n \xrightarrow{L^p} X} \arrow[r, Rightarrow, "p \geqslant q"] & \boxed{X_n \xrightarrow{L^q} X} \arrow[u, Rightarrow] &
\end{tikzcd}$$

### A.8.2 Converse Results on Modes of Convergence

1. *Convergence in distribution to a constant* [2, Page 24]: Let $c$ be a constant element of $E$. Then,

$$X_n \xrightarrow{d} c \quad \Rightarrow \quad X_n \xrightarrow{\mathbb{P}} c \, .$$

2. *Convergence in probability combined with uniform integrability* [28, Page 131]:
☞ 613 Suppose the numerical random variables $\{X_n\}$ are uniformly integrable.
Then, for $p \geqslant 1$,

$$X_n \xrightarrow{\mathbb{P}} X \quad \Rightarrow \quad X_n \xrightarrow{L^p} X \, .$$

This includes the case where $|X_n| \leqslant Y$ for all $n$ with $\mathbb{E}Y < \infty$ (dominated convergence).

3. *Continuity theorem* [2, Page 30]: Let $h : E \to E'$ be a measurable function, with $(E', \mathcal{E}')$ a measurable space and $E'$ equipped with metric $\varrho'$. Let $D_h \in \mathcal{E}$ be the set of discontinuities of $h$. If $\mathbb{P}(X \in D_h) = 0$ (in particular, when $h$ is continuous), then

$$X_n \xrightarrow{d} X \quad \Rightarrow \quad h(X_n) \xrightarrow{d} h(X) \, .$$

A special case is **Slutsky's theorem**: if $E = \mathbb{R}^2$ and $E' = \mathbb{R}$, then we have $(X_n, Y_n) \xrightarrow{d} (X, c)$, where $c \in \mathbb{R}$ is a constant, implies $h(X_n, Y_n) \xrightarrow{d} h(X, c)$ for all continuous functions $h : \mathbb{R}^2 \to \mathbb{R}$.

4. *Finite expectation of infinite series*: Let $X_n \geqslant 0$. If the infinite series $\sum_n X_n$ has finite expectation, then $X_n \xrightarrow{\text{a.s.}} 0$.

5. *Skorohod representation* [11, Page 271]: If $X_n \xrightarrow{d} X$ with corresponding distributions $P_{X_n}$ and $P_X$, then there exist random variables $\widetilde{X}_1, \widetilde{X}_2, \dots, \widetilde{X}$

in $(E, \mathcal{E})$ with distributions $P_{X_n}$ for each $n$ and $P_X$, respectively, such that $\widetilde{X}_n \xrightarrow{\text{a.s.}} \widetilde{X}$.

6. *Monotone convergence*: Suppose $\mathbb{E}X_n$ exists for some $n$. Then, for any $p \geqslant 1$,

$$X_n \overset{a.s.}{\nearrow} X \quad \Rightarrow \quad X_n \overset{L^p}{\nearrow} X \,.$$

### A.8.3 Law of Large Numbers and Central Limit Theorem

We briefly discuss two of the main results in probability: the law of large numbers and the central limit theorem. Both are associated with sums of independent random variables. For details, see, for example, [3, Pages 85, 357, and 385].

Let $X_1, X_2, \ldots$ be iid random variables with expectation $\mu$. The law of large numbers states that the sample average $(X_1 + \cdots + X_n)/n$ is close to $\mu$ for large $n$.

**Theorem A.8.1 (Strong Law of Large Numbers)** *Let $X_1, \ldots, X_n$ be iid with expectation $\mu$. Then,*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{\text{a.s.}} \mu \quad \textit{as} \;\; n \to \infty \,.$$

The central limit theorem describes the limiting distribution of the sum $S_n = X_1 + \cdots + X_n$. Loosely, it states that the random sum $S_n$ has a distribution that is approximately normal (Gaussian) when $n$ is large. The more precise statement is given next.

**Theorem A.8.2 (Central Limit Theorem)** *Let $X_1, \ldots, X_n$ be iid with expectation $\mu$ and variance $\sigma^2 < \infty$. Then,*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Y \sim \mathsf{N}(0,1) \quad \textit{as} \;\; n \to \infty \,.$$

In other words, for large $n$ the random sum $S_n$ has a distribution that is approximately normal with expectation $n\mu$ and variance $n\sigma^2$. Under the extra condition that $\mathbb{E}|X - \mu|^3 < \infty$, precise error bounds can be found on the standardized cdf of $S_n$. Below, $\Phi$ is the cdf of $Y \sim \mathsf{N}(0,1)$.

**Theorem A.8.3 (Berry–Esséen)** *Let $X_1, \ldots, X_n$ be iid with expectation $\mu$ and variance $\sigma^2 < \infty$. Then, for all $n$,*

$$\sup_x \left| \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) - \Phi(x) \right| \leqslant K \frac{\mathbb{E}|X_1 - \mu|^3}{\sqrt{n}\,\sigma^3} \,,$$

*for some constant $K > 0$ that does not depend on $n$ or the distribution of $X_1$.*

For a proof, see, for example, [5, Page 224]. The smallest constant $K$ found to date is $K = 0.7056$, see [27].

**Theorem A.8.4 (Multivariate Central Limit Theorem)** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid random vectors with expectation vector $\boldsymbol{\mu}$ and finite covariance matrix $\Sigma$. Define $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$. Then,*

$$\frac{\mathbf{S}_n - n\boldsymbol{\mu}}{\sqrt{n}} \xrightarrow{d} \mathbf{Y} \sim \mathsf{N}(\mathbf{0}, \Sigma) \quad \textit{as} \;\; n \to \infty \,.$$

## A.9 STOCHASTIC PROCESSES

A **stochastic process** or **random process** is a collection of random variables $\{X_t, t \in \mathscr{T}\}$ on a probability space $(\Omega, \mathcal{H}, \mathbb{P})$, where $\mathscr{T}$ is any index set. The set $E$ of possible values for $X_t$ (assuming this is independent of $t$) is called the **state space** of the process. The index set $\mathscr{T}$ is often taken to be a countable or continuous subset of $\mathbb{R}$, and so a stochastic process is often thought of as a random variable evolving through time, with $X_t$ representing the state of the process at time $t$.

The distribution of a stochastic process $X = \{X_t, t \in \mathscr{T}\}$, with $\mathscr{T} \subseteq \mathbb{R}$, is completely determined by its finite-dimensional distributions; that is, the distributions of the random vectors $(X_{t_1}, \ldots, X_{t_n})$ for any choice of $n$ and $t_1, \ldots, t_n$. However, the finite-dimensional distributions do not completely determine the sample path behavior of a stochastic process; see, for example, [3, Page 308]. Hence, questions of continuity and differentiability cannot be answered by examining the finite-dimensional distributions alone. Processes that share the same finite-dimensional distributions are called **versions** of each other. If, in addition, the processes share the same probability space, then they are called **modifications** of each other. For a consistent system of finite-dimensional distributions it is always possible to choose a version of the stochastic process that (almost surely) has separable paths [3, Pages 526–527]. A path $\{x_t, t \in \mathscr{T}\}$ is said to be **separable** if there exists a countable dense subset $\mathscr{D}$ of $\mathscr{T}$, such that for each $t \in \mathscr{T}$ there exists a sequence $t_1, t_2, \cdots \in \mathscr{D}$ with $t_n \to t$ and $x_{t_n} \to x_t$. The sample path behavior of a separable process is determined by its finite-dimensional distributions. We will assume henceforth that we are dealing with the separable versions of stochastic processes.

### ■ EXAMPLE A.5 (Bernoulli Process)

A basic example of a stochastic process is any collection $\{X_1, X_2, \ldots\}$ of iid random variables. When $X_t \sim_{\text{iid}} \mathsf{Ber}(p)$ for $t = 1, 2, \ldots$ the process is called a **Bernoulli process**. Here the state space is $E = \{0, 1\}$ and the index set is $\mathscr{T} = \{1, 2, \ldots\}$. The process models the random experiment where a biased coin is tossed indefinitely. The beginning of a typical **sample path** of the process for $p = 0.5$ is given in Figure A.6.

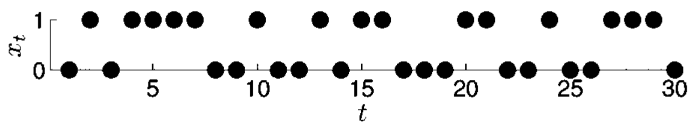

**Figure A.6** A typical sample path for a Bernoulli process with $p = 0.5$.

The description and study of real-valued stochastic processes that evolve over time are facilitated by the following notions. In all cases $\mathscr{T}$ is assumed to be one of $\mathbb{N}, \mathbb{Z}, \mathbb{R}_+$, or $\mathbb{R}$.

A collection $\{\mathcal{H}_t\} = \{\mathcal{H}_t, t \in \mathscr{T}\}$ of $\sigma$-algebras of events, with the property that $\mathcal{H}_t \subseteq \mathcal{H}_{t+s}$ for any $s \geqslant 0$ and $t \in \mathscr{T}$, is called a **filtration** or **history**. A filtration is called **right-continuous** if $\mathcal{H}_t = \mathcal{H}_{t+} \overset{\text{def}}{=} \cap_{s>t} \mathcal{H}_s$ for all $t$. A filtration can be thought of as an increasing flow of information about some random phenomenon. A stochastic process $\{X_t, t \in \mathscr{T}\}$ is called **adapted** to a filtration $\{\mathcal{H}_t\}$ if $X_t$ is $\{\mathcal{H}_t\}$-measurable, for every $t \in \mathscr{T}$; that is, $X_t \in \mathcal{H}_t$ for all $t$. Intuitively, $\mathcal{H}_s, s \leqslant t$ contains the complete history of the process up to time $t$. A random variable $\tau \in \mathscr{T}$ is said to be a **stopping time** with respect to $\{\mathcal{H}_t\}$ if for each $t \in \mathscr{T}$ the event $\{\tau \leqslant t\}$ lies in $\mathcal{H}_t$. Intuitively, $\tau$ is a stopping time if one can decide if it has occurred by time $t$ on the basis of the information (contained in $\mathcal{H}_t$) up until time $t$.

### ■ EXAMPLE A.6 (Bernoulli Process Continued)

A rich variety of stochastic processes can be derived from a Bernoulli process $\{X_t, t = 1, 2, \ldots\}$. For example, define $S_0 = 0$ and $S_t = S_{t-1} + X_t$, $t = 1, 2, \ldots$. Process $\{S_t\}$ is an example of a **random walk** process. Let $\mathcal{H}_t$ be the history of the Bernoulli process up until time $t$. Note that $\{S_t\}$ is adapted to the filtration $\{\mathcal{H}_t\}$, because all information regarding $S_1, \ldots, S_t$ can be obtained from $X_1, \ldots, X_t$ and vice versa. Let $\tau_n$ be the first time that $\{S_t\}$ crosses level $n$, that is, $\tau_n = \inf\{t : S_t \geqslant n\}$. Then, $\tau_n$ is a stopping time with respect to $\{\mathcal{H}_t\}$, because the occurrence of $\{\tau_n \leqslant t\}$ can be decided upon using information about $X_1, \ldots, X_t$ only.

## A.9.1 Gaussian Property

A real-valued stochastic process $\{X_t, t \in \mathscr{T}\}$ is said to be **Gaussian** if all its finite-dimensional distributions are Gaussian (normal); that is, if the vector $(X_{t_1}, \ldots, X_{t_n})$ is multidimensional Gaussian for any choice of $n$ and $t_1, \ldots, t_n \in \mathscr{T}$, or equivalently, if any linear combination $\sum_{i=1}^{n} b_i X_{t_i}$ has a Gaussian distribution. Gaussian processes can thus be thought of as generalizations of Gaussian random ☞ 143
vectors.

The probability distribution of a Gaussian process is determined completely by its **expectation function**

$$\mu_t = \mathbb{E}X_t, \quad t \in \mathscr{T}$$

and **covariance function**

$$\sigma_{s,t} = \mathrm{Cov}(X_s, X_t), \quad s, t \in \mathscr{T} .$$

A zero-mean Gaussian process is one for which $\mu_t = 0$ for all $t$. The generation of Gaussian processes is discussed in Section 5.1. ☞ 154

### ■ EXAMPLE A.7 (Wiener Process)

The **Wiener process** $\{W_t, t \geqslant 0\}$ can be defined as a zero-mean Gaussian process with covariance function

$$\sigma_{s,t} = \min\{s, t\} , \quad s, t \geqslant 0 .$$

It forms the basis of a great variety of other stochastic processes; see Chapter 5 and Sections A.12–A.13. The Wiener process has many interesting properties and characterizations, which are further discussed in Section 5.5. ☞ 177

### A.9.2 Markov Property

A stochastic process $\{X_t, t \in \mathscr{T}\}$ on $(\Omega, \mathcal{H}, \mathbb{P})$, with index set $\mathscr{T} \subseteq \mathbb{R}$ and state space $E$ (equipped with a $\sigma$-algebra $\mathcal{E}$), is said to be a **Markov process** if for every $s \geqslant 0$, $t \in \mathscr{T}$, and $A \in \mathcal{E}$ it satisfies the **Markov property**:

$$\mathbb{P}(X_{t+s} \in A \,|\, \mathcal{H}_t) = \mathbb{P}(X_{t+s} \in A \,|\, X_t)\,, \tag{A.34}$$

where $\mathcal{H}_t$ is the history of the process up until time $t$. The Markov process is said to be **time-homogeneous** if the conditional probability $P_s(x, A) = \mathbb{P}(X_{t+s} \in A \,|\, X_t = x)$ does not depend on $t$ for any fixed $s$. The function $P_s$ is called the ($s$-step) **transition kernel** of the Markov process. When (A.34) holds for any stopping time $\tau$ instead of a fixed $t$, then $\{X_t\}$ is said to have the **strong Markov property**.

The Markov property can be expressed as

$$(X_{t+s} \,|\, X_u, u \leqslant t) \quad \sim \quad (X_{t+s} \,|\, X_t)\,, \tag{A.35}$$

which emphasizes that the conditional future distributions of the sample path given the entire sample path history are the same as those given only the present state. In other words, for a Markov process the conditional distribution of the "future" variable $X_{t+s}$ given the entire past of the process $\{X_u, u \leqslant t\}$ is the same as the conditional distribution of $X_{t+s}$ given only the "present" $X_t$.

We assume from now on that the Markov process is time-homogeneous, unless otherwise specified. Markov processes come in many different varieties, depending on the choice of index set $\mathscr{T}$ and state space $E$. In most cases of practical interest $\mathscr{T} = \mathbb{N}$ or $\mathbb{R}_+$ and $E \subseteq \mathbb{R}^n$. In addition, in many cases $P_t$ is of the form

$$P_t(x, A) = \int_{y \in A} p_t(x, y)\,\mathrm{d}y \quad \text{or} \quad P_t(x, A) = \sum_{y \in A} p_t(x, y)\,, \tag{A.36}$$

in the continuous and discrete case, respectively. Here, $p_t(x, y)$ is the **transition kernel density**. In this case the finite-dimensional distributions of the Markov process (and hence the distribution of the entire process) are determined by the family of transition kernels $\{P_t, t \geqslant 0\}$ and the distribution of $X_0$ — the **initial distribution** of the Markov process. Namely, by the product rule (A.21) and the Markov property the joint probability density $f$ of any random vector $(X_0, X_{t_1}, \ldots, X_{t_n})$ satisfies

$$f(x_0, x_1, \ldots, x_n) = f_{X_0}(x_0)\, p_{t_1}(x_0, x_1)\, p_{t_2 - t_1}(x_1, x_2) \cdots p_{t_n - t_{n-1}}(x_{n-1}, x_n)\,,$$

where $f_{X_0}$ is the density of $X_0$. The kernel $P_t$ can be viewed as a linear operator $f \mapsto P_t f$ acting on suitable functions $f$, such that

$$P_t f(x) \stackrel{\text{def}}{=} \mathbb{E}^x f(X_t) = \int P_t(x, \mathrm{d}y) f(y)\,.$$

Here $\mathbb{E}^x$ denotes the expectation operator under which the process starts in $x$ at time 0. An important property of $\{P_t, t \geqslant 0\}$ is the **semigroup** property:

$$P_{s+t} = P_s\, P_t \quad \text{for all } s, t \geqslant 0\,. \tag{A.37}$$

These are the **Chapman–Kolmogorov** equations.

Sections A.10 and A.11 discuss discrete-time Markov processes, often called **Markov chains**, and continuous-time Markov processes in greater detail.

### A.9.3 Martingale Property

A **martingale** is a real-valued stochastic process $X = \{X_t, t \in \mathscr{T}\}$, with $\mathscr{T} \subseteq \mathbb{R}$, such that:

1. $X$ is adapted to a filtration $\{\mathcal{H}_t\}$.

2. $\mathbb{E}|X_t| < \infty$ for all $t \in \mathscr{T}$.

3. For any $s \leqslant t \in \mathscr{T}$,
$$\mathbb{E}[X_t \,|\, \mathcal{H}_s] = X_s\,, \quad \text{a.s.} \tag{A.38}$$

The state $X_t$ of the process can be interpreted as the fortune at time $t$ of a gambler playing a game. In this context a martingale can be thought of as a "fair game", in the sense that the gambler's fortune in the future is expected to be the same as the gambler's current fortune, given all the past and present information on the game. In some cases it is important to stress the filtration $\{\mathcal{H}_t\}$ and probability measure $\mathbb{P}$ under which the above martingale conditions hold.

A process $X$ is called a **submartingale** if (A.38) holds with "=" replaced by "$\geqslant$". An **$L^p$-(sub)martingale** is a (sub)martingale for which $\mathbb{E}|X_t|^p < \infty$ for all $t$. One usually distinguishes between **discrete-time** ($\mathscr{T} = \mathbb{N}$ or $\mathbb{Z}$) and **continuous-time** ($\mathscr{T} = \mathbb{R}_+$ or $\mathbb{R}$) martingales. The properties of continuous-time martingales are similar to those of the discrete-time equivalents, but often additional regularity conditions are required. We list a number of properties of martingales. For proofs, see [7].

1. *Sample path regularity*: Let $X = \{X_t, t \geqslant 0\}$ be a submartingale such that $t \mapsto \mathbb{E}X_t$ is continuous. Then, $X$ has a modification that has right-continuous and left-limited paths (this is automatically so if $X$ is a martingale).

2. *Maximum bound*: Let $\{X_t, t = 0, 1, 2, \ldots\}$ be an $L^p$-martingale for some $p \geqslant 1$. Then,
$$\mathbb{P}\Big(\max_{0 \leqslant t \leqslant n} |X_t| \geqslant x\Big) \leqslant \frac{\mathbb{E}|X_n|^p}{x^p}, \quad x \geqslant 0\,.$$

3. *Convergence*: Let the process $X = \{X_t, t = 0, 1, 2, \ldots\}$ be a (sub)martingale. If $\sup_n \mathbb{E}X_n^+ < \infty$, where $x^+ = \max\{x, 0\}$, then $X$ converges almost surely to an integrable random variable $X_\infty$.

4. *Optional sampling*: Let $X = \{X_t, t \geqslant 0\}$ be a (sub)martingale and $\tau_1, \tau_2, \ldots$ be a sequence of stopping times such that $\tau_i \leqslant K_i$ for some deterministic sequence $K_1, K_2, \ldots < \infty$. Then, $\{X_{\tau_i}, i = 1, 2, \ldots\}$ is a (sub)martingale with respect to filtration $\{\mathcal{H}_{\tau_i}\}$.

5. *Optional stopping*: Let the process $X = \{X_t, t \geqslant 0\}$ be a martingale and $\tau$ a finite stopping time. If $X$ is uniformly integrable, then $X_\tau = \mathbb{E}[X_\infty \,|\, \mathcal{H}_\tau]$ and $\mathbb{E}X_\tau = \mathbb{E}X_0$.

6. *Criterion for martingale:* Let $\{X_t, t \geqslant 0\}$ be a process such that $\mathbb{E}|X_\tau| < \infty$ and $\mathbb{E}X_\tau = \mathbb{E}X_0$ for every bounded stopping time $\tau \leqslant K < \infty$. Then, $X$ is a martingale.

7. *Submartingale implying martingale:* Let $X = \{X_t, t \geqslant 0\}$ be a submartingale on $t \in [0, T]$. If $\mathbb{E}X_T = \mathbb{E}X_0$, then $X$ is a martingale on $t \in [0, T]$.

8. *Martingale representation:* Let $\{X_t, 0 \leqslant t \leqslant T\}$ be a square-integrable martingale. Then there exists a unique process $\{\phi_t\}$ adapted to $\{\mathcal{H}_t\}$ such that:

   (a) $\mathbb{E} \int_0^T \phi_t^2 \, \mathrm{d}t < \infty$;

   (b) $X_t = X_0 + \int_0^t \phi_s \, \mathrm{d}W_s$, $t \in [0, T]$, where $\{W_t, t \geqslant 0\}$ is a Wiener process adapted to $\{\mathcal{H}_t\}$.

### A.9.4 Regenerative Property

A real-valued stochastic process $X = \{X_t, t \geqslant 0\}$ is said to be **regenerative** if there exist times $T_0 \leqslant T_1 < T_2 < T_3 < \ldots$ of the form $T_n = A_1 + \cdots + A_n$, $n = 1, 2, \ldots$, where the $\{A_i\}$ are iid, such that conditional on $X_s, s \leqslant T_n$ the process $\{X_{T_n+t}, t \geqslant 0\}$ has the same distribution as $\{X_{T_0+t}, t \geqslant 0\}$. In other words, a regenerative process "regenerates" itself at times $T_0, T_1, \ldots$. That is, given the history of the process up to time $T_n$, the process after $T_n$ behaves probabilistically as if it has started afresh. The $\{T_n\}$ are called **regeneration times**. When $T_0 = 0$, the process is called **pure**; otherwise, it is called **delayed**.

### ■ EXAMPLE A.8 ($M/M/1$ Queue)

The $M/M/1$ **queueing system** describes a service facility where customers arrive at certain random times and are served by a single server. Arriving customers who find the server busy wait in the queue. Customers are served in the order in which they arrive. The interarrival times are iid exponential random variables with rates $\lambda$, and the service times of customers are iid exponential random variables with rates $\mu$. Finally, the service times are independent of the interarrival times. Assume that at $T_0 = 0$ the system is empty and that mean service time is smaller than the mean interarrival time. Let $X_t$ be the number of customers in the system at time $t$. Then $\{X_t, t \geqslant 0\}$ is a regenerative process. Namely, let $T_1$ be the first time the system becomes empty again after a service completion. The probabilistic behavior of the process $\{X_t\}$ from $T_1$ onwards is exactly the same as from $t = 0$ onwards, even if we knew the complete history up to time $T_1$. Let $T_2$ be the next time the system becomes empty after a service completion, and so on. Figure A.7 shows a realization of this process.

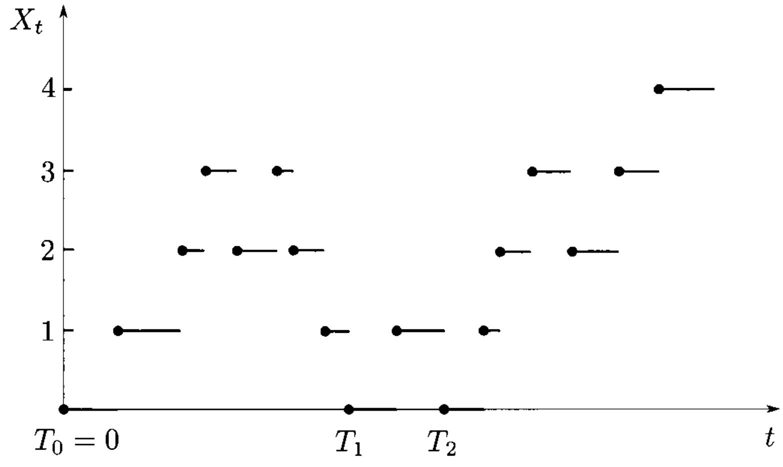

**Figure A.7** The number of customers in an $M/M/1$ queue as a function of time.

The process $\{T_n\}$ of renewal times forms a so-called **renewal process**; the corresponding $\{A_n\}$ are called **cycle lengths**. The following main property of regenerative processes is derived from the properties of renewal processes; see, for example, [4, Chapter 9]. A random variable $A$ is said to have a **lattice** distribution if $A$ takes values in the lattice $\{a + b\,n, n \in \mathbb{Z}\}$ for some values of $a$ and $b$ ($b \neq 0$); $b$ is called the **period**.

**Theorem A.9.1 (Regeneration Theorem)** *Let $\{X_t\}$ be a continuous-time regenerative process with right-continuous paths and nonlattice distribution of the cycle length with expectation $\mu = \mathbb{E}A_1 < \infty$. Then, $X_t$ converges in distribution to a random variable $X$, such that for all $f$*

$$\mathbb{E}f(X) = \frac{1}{\mu}\mathbb{E}\int_{T_0}^{T_1} f(X_s)\,\mathrm{d}s\,, \tag{A.39}$$

*provided that the expectation exists.*

*Let $\{X_t\}$ be a discrete-time regenerative process with cycle length distribution of period $b = 1$ and expectation $\mu = \mathbb{E}A_1 < \infty$. Then, $X_n$ converges in distribution to a random variable $X$, such that for all $f$*

$$\mathbb{E}f(X) = \frac{1}{\mu}\mathbb{E}\sum_{k=T_0}^{T_1-1} f(X_k)\,, \tag{A.40}$$

*provided that the expectation exists.*

In other words, if $G_t$ denotes the cdf of $X_t$ ($G_t(x) = \mathbb{P}(X_t \leqslant x)$), then under the mild conditions above, there *exists* a continuous cdf $G$ such that $\lim_{t\to\infty} G_t(x) = G(x)$ for all $x$.

Often $G_t$ is difficult to calculate, but $G$ is usually much easier to find, via equation (A.39) or (A.40). Moreover, with the existence of $G$ guaranteed, we can now give a precise meaning to the behavior of the stochastic process "in the stationary situation" or "in equilibrium".

■ **EXAMPLE A.9 ($M/M/1$ Queue Continued)**

As in Example A.8, let $X_t$ denote the number of customers in an $M/M/1$ queueing system at time $t$. When the arrival rate is smaller than the service rate, $\{X_t\}$ is a regenerative process, and hence $X_t$ converges in distribution to a random variable $X$ that can be interpreted as the number of customers in the system "in equilibrium" or far into the future. Similarly, the expected steady-state number of customers in the stationary situation simply refers to the expectation of $X$.

### A.9.5 Stationarity and Reversibility

A stochastic process $\{X_t, t \in \mathscr{T}\}$ is said to be **strongly stationary** if the distributions of the random vectors $(X_{t_1}, \dots, X_{t_n})$ and $(X_{t_1+s}, \dots, X_{t_n+s})$ are the same for any choice of $n$ and $s, t_1, \dots, t_n \in \mathscr{T}$.

A stochastic process $\{X_t, t \in \mathscr{T}\}$ is said to be **weakly stationary** if both the expectation function $\{\mathbb{E}X_t\}$ and covariance function $\{\mathrm{Cov}(X_t, X_{t+s})\}$ do not depend on $t$. The function $R(s) = \mathrm{Cov}(X_t, X_{t+s})$ is then called the **autocovariance function**.

In other words, the distribution of a strongly stationary process is invariant under time shifts (or space shifts in cases where $\mathscr{T}$ is a spatial index set). For weakly stationary processes the covariance function is invariant under time shifts. A strongly stationary process is weakly stationary whenever its mean and covariance function exist. In particular, this is the case when $\mathbb{E}X_t^2 < \infty$, $t \in \mathscr{T}$. However, a weakly stationary process is not necessarily strongly stationary. A notable exception to this are Gaussian processes (see Section A.9.1), as their finite-dimensional distributions depend only on the corresponding means and covariances.

A strongly stationary stochastic process $\{X_t\}$ with index set $\mathbb{Z}$ or $\mathbb{R}$ is said to be **reversible** if, for any positive integer $n$ and for all $t_1, \ldots, t_n$, the vector $(X_{t_1}, \ldots, X_{t_n})$ has the same distribution as $(X_{-t_1}, \ldots, X_{-t_n})$. One way to visualize reversible processes is to imagine that we have taken a video of the stochastic process which we may run in forward and reverse time. If we cannot detect whether the video is running forward or backward, the process is reversible.

## A.10 MARKOV CHAINS

A Markov process (see Section A.9.2) with a countable index set $\mathscr{T}$ is called a **Markov chain**. Below, we assume that the index set is either $\mathbb{N}$ or $\mathbb{Z}$ and that the chain is time-homogeneous. Generating realizations of a Markov chain is discussed in Section 5.2. ☞ 162

Recall from Section A.9.2 that the transition kernel $P_t(x, A)$ of a general time-homogeneous Markov process gives the probability that starting from $x$ the chain ends up in set $A$ after $t$ discrete time steps. Of particular importance for Markov chains is the one-step transition kernel $P_1$. If the state space $E$ is countable, say $E = \mathbb{N}$, we can write its (discrete) density as

$$p(x,y) = P_1(x, \{y\}) = \mathbb{P}(X_{t+1} = y \mid X_t = x), \quad x, y \in E, \quad t \in \mathbb{N}\,. \tag{A.41}$$

We can arrange these one-step transition probabilities in a one-step **transition matrix** $P$ with $(x,y)$-th entry given by $p(x,y)$. Similarly, $P_t$ is represented by the $t$-step transition matrix with $(x,y)$-th element $p_t(x,y) = P_t(x, \{y\})$. Note that the elements of $P_t$ in every row are nonnegative and sum up to unity. Such a matrix is called a **stochastic** matrix. If additionally every column sums to unity, then the matrix is called **doubly stochastic**.

By the Chapman–Kolmogorov equations (A.37), the $t$-step transition matrix is in fact equal to the $t$-th power of $P$; that is, $P_t = P^t$. It follows that if $\boldsymbol{\pi}_t = (\mathbb{P}(X_t = k), k \in E)$ is the row vector representing the probability distribution of $X_t$, then

$$\boldsymbol{\pi}_t = \boldsymbol{\pi}_0 P^t \quad \text{for all } t = 0, 1, \ldots, \tag{A.42}$$

where $P^0$ is the identity matrix.

When $E$ is nondenumerable, for example $E = \mathbb{R}$, and $P_t$ has a density $p_t$ as in (A.36), the one-step transition matrix is replaced by the one-step transition density $p(x,y) = p_1(x,y)$. The Chapman–Kolmogorov equations for the transition densities become

$$p_{t+s}(x,y) = \int_E p_s(x,z)\, p_t(z,y)\, \mathrm{d}z, \quad s,t \in \mathbb{N}, \quad x,y \in E\,. \tag{A.43}$$

A convenient way to describe a discrete-state Markov chain $X$ is through its **transition graph**. States are indicated by the nodes of the graph (without the weight labels), and a strictly positive (> 0) transition probability $p(x, y)$ from state $x$ to $y$ is indicated by an arrow from $x$ to $y$ with weight $p(x, y)$. An example of a transition graph is given in Figure A.8.

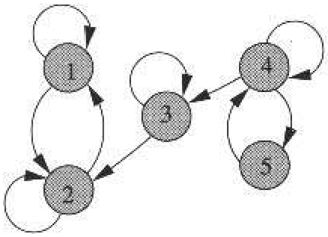

**Figure A.8** A transition graph of a discrete-state Markov chain.

### A.10.1 Classification of States

Let $X = \{X_t, t = 0, 1, \dots\}$ be a time-homogeneous Markov chain with state space $E$. Let $x$ and $y$ be arbitrary states in $E$. Let $T$ denote the time that the chain first visits state $y$, or first returns to $y$ if it started there; and let $N_y$ denote the total number of visits to $y$ from time 0 onwards. We write $\mathbb{P}^y(A)$ for $\mathbb{P}(A \mid X_0 = y)$ for any event $A$. We denote the corresponding expectation operator by $\mathbb{E}^y$. The states of a Markov chain are typically classified as follows.

1. A state $y$ is called a **recurrent** state if $\mathbb{P}^y(T < \infty) = 1$; otherwise, it is called **transient**. A recurrent state $y$ is called **positive recurrent** if $\mathbb{E}^y T < \infty$; otherwise, it is called **null-recurrent**.

2. A state $y$ is said to be **periodic with period $\delta$**, if $\delta \geqslant 2$ is the largest integer for which $\mathbb{P}^y(T = n\delta, \text{ for some } n \geqslant 1) = 1$. If $\delta = 1$, the state is said to be **aperiodic**.

3. If $p_t(x, y) > 0$ for some $t \geqslant 0$, then $x$ is said to **lead to** $y$ — written as $x \to y$. If $x \to y$ and $y \to x$, then $x$ and $y$ are said to **communicate** — written as $x \leftrightarrow y$. A set of states $C \subseteq E$ is called a **communicating class** if, for any pair $x, y \in C$, $x \leftrightarrow y$, and further that for every $x \in C$ there is no $y \in E \setminus C$ such that $x \leftrightarrow y$. If $E$ is the only communicating class, the Markov chain is said to be **irreducible**.

4. A set of states $A \subseteq E$ such that $\sum_{y \in A} p(x, y) = 1$ for all $x \in A$ is called a **closed** set. A state $x$ is called an **absorbing** state if $\{x\}$ is closed.

Recurrence and transience are class properties; that is, the elements in each communicating class are either all recurrent or all transient. Figure A.8 shows the transition graph of a Markov chain with three communicating classes.

### A.10.2 Limiting Behavior

The limiting or steady-state behavior of Markov chains as $t \to \infty$ is of considerable interest and importance, and is often simpler to describe and analyze than the transient behavior of the chain for fixed $t$.

For simplicity, assume that the state space $E$ is countable. Then, a Markov chain $\{X_t\}$ is a discrete-time regenerative process, where possible renewal times are the times when the process returns to a specific state. Irreducibility and aperiodicity ensure, via Theorem A.9.1, that

$$\lim_{t \to \infty} p_t(x, y) = \pi(y) , \tag{A.44}$$

for some $\pi(y) \in [0, 1]$. Moreover, $\pi(y) > 0$ if $y$ is positive recurrent and $\pi(y) = 0$ otherwise. The intuitive reason behind this result is that the process "forgets" where it was initially if it goes on long enough. Thus, provided that $\pi(y) \geqslant 0$ and $\sum_y \pi(y) = 1$, the numbers $\{\pi(y), y \in E\}$ form the **limiting distribution** of the Markov chain. Note that these conditions are not always satisfied. For example, they are clearly not satisfied if the Markov chain is transient, and they may not be satisfied even if the chain is recurrent (namely when the states are null-recurrent). When $E = \{0, 1, 2, \ldots\}$, then the limiting distribution is usually identified with the row vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$. The following is proved, for example, in [4].

**Theorem A.10.1 (Limiting Distribution)** *For an irreducible aperiodic Markov chain with transition matrix $P$, if the limiting distribution $\boldsymbol{\pi}$ exists, then $\boldsymbol{\pi}$ is uniquely determined by the solution of the constrained system of equations*

$$\boldsymbol{\pi} = \boldsymbol{\pi} P, \quad \sum_{y \in E} \pi_y = 1, \quad \pi_y \geqslant 0 \quad \textit{for all } y \in E . \tag{A.45}$$

*In fact, the solution of* (A.45) *will automatically be strictly positive* ($\pi_y > 0$). *Conversely, if there exists a row vector $\boldsymbol{\pi}$ satisfying* (A.45), *then $\boldsymbol{\pi}$ is the limiting distribution of the Markov chain. In addition, $\pi_y > 0$ for all $y$, and all states are positive recurrent.*

Let $X$ be a Markov chain with limiting distribution $\boldsymbol{\pi}$. Suppose $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$. Then, combining (A.42) and (A.45), we have $\boldsymbol{\pi}_t = \boldsymbol{\pi}$ . Thus, if the initial distribution of the Markov chain is equal to the limiting distribution, then the distribution of $X_t$ is the same for all $t$ and is given by this limiting distribution. For any Markov chain, any $\boldsymbol{\pi}$ which satisfies (A.45) is called a **stationary distribution**, because using $\boldsymbol{\pi}$ as an initial distribution renders the Markov chain a stationary process.

Noting that $\sum_y p(x, y) = 1$, we can rewrite (A.45) as the system of equations

$$\sum_y \pi(x)\, p(x, y) = \sum_y \pi(y)\, p(y, x) \quad \text{for all } x \in E . \tag{A.46}$$

These are called the **global balance equations**. We can interpret (A.45) as the statement that the "probability flux" out of $x$ is balanced with the probability flux into $x$. An important generalization, which follows directly from (A.46), states that the same balancing of probability fluxes holds for an arbitrary set $A$. That is, for every set $A \subseteq E$ of states we have

$$\sum_{x \in A} \sum_{y \notin A} \pi(x)\, p(x, y) = \sum_{x \in A} \sum_{y \notin A} \pi(y)\, p(y, x) . \tag{A.47}$$

### A.10.3 Reversibility

A good way to think of the global balance equations (A.46) is that they balance the probability flux out of each state $x$ with the probability flux into state $x$. For *reversible* (see Section A.9.5) Markov chains a much stronger form of balance equations holds, where the probability flux from state $x$ to state $y$ is balanced with that from state $y$ to state $x$. The following theorem is proved in [17, 20].

**Theorem A.10.2 (Reversible Markov Chain)** *A stationary Markov chain is reversible if and only if there exists a collection of positive numbers $\{\pi(x),\ x \in E\}$, summing to unity that satisfy the* **detailed (or local) balance equations**

$$\pi(x)\, p(x,y) = \pi(y)\, p(y,x)\ , \quad x, y \in E\ . \tag{A.48}$$

*Whenever there exists such a collection $\{\pi(x)\}$, it is the stationary distribution of the process.*

The following gives a simple criterion for reversibility based on the transition probabilities. A proof can be found in [17, Page 21].

**Theorem A.10.3 (Kolmogorov's Criterion)** *A stationary Markov chain is reversible if and only if its transition probabilities satisfy*

$$p(x_1,x_2)\,p(x_2,x_3)\cdots p(x_{n-1},x_n)\,p(x_n,x_1) = p(x_1,x_n)\,p(x_n,x_{n-1})\cdots p(x_2,x_1) \tag{A.49}$$

*for all finite loops of states $x_1, \ldots, x_n, x_1$.*

The idea is quite intuitive: if the process in forward time is more likely to traverse a certain closed loop in one direction than in the opposite direction, then in backward time it will exhibit the opposite behavior, and hence we have a criterion for detecting the direction of time. If such "looping" behavior does not occur, the process must be reversible.

## A.11 MARKOV JUMP PROCESSES

A **Markov jump process** is a Markov process (see Section A.9.2) with a continuous index set and a discrete (that is, countable) state space $E$. Generating realizations of a Markov jump process is discussed in Section 5.3. For simplicity ☞ 166 we assume that the Markov jump process is time-homogeneous and that the index set is either $\mathbb{R}$ or $\mathbb{R}_+$. Let $p_t(x,y) = P_t(x,\{y\}) = \mathbb{P}(X_t = y \,|\, X_0 = x)$ denote the **transition probability** from $x$ to $y$ in $t \geqslant 0$ time units. Similar to a Markov chain with a discrete state space, we can arrange the transition probabilities into a matrix $(p_t(x,y))$. With a slight abuse of notation we will also write this matrix as $P_t$. We will call the family $\{P_t, t \geqslant 0\}$, or $P_t$ viewed as a function $t$, a **transition function**. It is said to be **standard** if $\lim_{t\downarrow 0} P_t = I$ (the identity matrix) and **honest** if $P_t\mathbf{1} = \mathbf{1}$ for all $t$, where $\mathbf{1}$ is a column vector of ones. We will consider only standard transition functions.

The analogue of the one-step transition matrix for Markov chains is the ***Q*-matrix** defined as

$$Q = P_0' = \lim_{t\downarrow 0} \frac{P_t - I}{t}\ . \tag{A.50}$$

The $(x,y)$-th entry $(x \neq y)$ of $Q$, denoted $q(x,y)$, is called the **transition rate** from $x$ to $y$. The $x$-th diagonal entry, $q(x,x)$, is written as $-q_x$. It can be shown [1] that

(a) $0 \leqslant q(x,y) \leqslant \infty$, $x \neq y$,

(b) $\sum_{y \neq x} q(x,y) \leqslant q_x$.

A state $x$ is said to be **stable** if $q_x < \infty$; and **instantaneous** if $q_x = \infty$. If $q_x = 0$ the state $x$ is called **absorbing**.

A Markov jump process is usually defined by specifying a matrix $Q$ that satisfies the properties (a) and (b) above. Such a matrix is again called a $Q$-matrix. It is said to be **stable** if all the states are stable, **uniformly bounded** if $\sup_x q_x < \infty$, and **conservative** if $Q\mathbf{1} = \mathbf{0}$. Finally, $Q$ is called **regular** if it is conservative and

$$Q\mathbf{z} = \lambda \mathbf{z}, \quad -1 \leqslant z_i \leqslant 1 \text{ for all } i \,,$$

has the unique trivial solution $\mathbf{z} = \mathbf{0}$ for all $\lambda > 0$. The following theorem is proved in [1].

**Theorem A.11.1 (Sample Path Behavior)** *For each stable and conservative $Q$-matrix there exists a Markov jump process $X$ whose paths are right-continuous step functions up to a certain random time $T_\infty$. Moreover, the sample path behavior up to $T_\infty$ can be described as follows:*

1. *Given its past, the probability that $X$ jumps from its current state $x$ to state $y$ is $K(x,y) = q(x,y)/q_x$.*
2. *The amount of time that $X$ spends in state $y$ has an $\mathsf{Exp}(q_y)$ distribution, independent of the past history.*

A typical sample path of $X$ is sketched in Figure A.9. The process jumps at times $T_1, T_2, \ldots$ to states $Y_1, Y_2, \ldots$, staying an exponentially distributed length of time in each state.

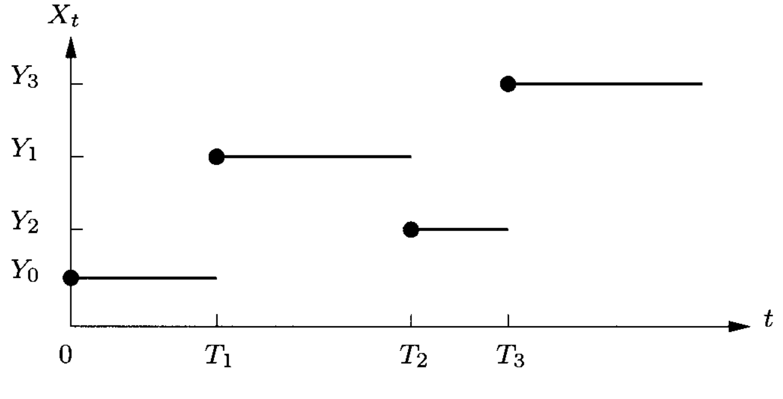

**Figure A.9** A sample path of a Markov jump process $\{X_t, t \geqslant 0\}$.

The first statement of Theorem A.11.1 implies that the process $\{Y_n, n \in \mathbb{N}\}$ is in fact a time-homogeneous Markov chain, with one-step transition matrix $K = (K(x,y))$. This Markov chain is called the **embedded Markov chain** or the **jump chain**.

A convenient way to describe a Markov jump process is through its **transition rate graph** (see, for example, Figure A.10). This is similar to a transition graph for Markov chains. The states are represented by the nodes of the graph, and a transition rate from state $x$ to $y$ is indicated by an arrow from $x$ to $y$ with weight $q(x, y)$.

Classification concepts such as irreducibility, communication, recurrence, and transience are defined in the same way as for a Markov chain; see Section A.10.1. Note, however, that there is no concept of periodicity for Markov jump processes.

### ■ EXAMPLE A.10 (Birth and Death Process)

A **birth and death process** is a Markov jump process with a transition rate graph of the form given in Figure A.10. Imagine that $X_t$ represents the total number of individuals in a population at time $t$. Jumps to the right correspond to "births", and jumps to the left to "deaths". The **birth rates** $\{b_i\}$ and the **death rates** $\{d_i\}$ may differ from state to state. Many applications of Markov chains involve processes of this kind.

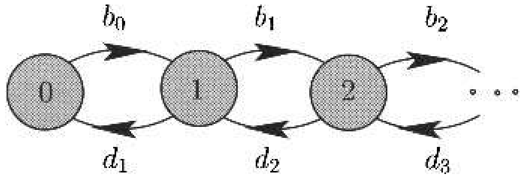

**Figure A.10** The transition rate graph of a birth and death process.

Note that the process jumps from one state to the next according to a Markov chain with transition probabilities $K_{0,1} = 1$, $K_{i,i+1} = b_i/(b_i + d_i)$ and $K_{i,i-1} = d_i/(b_i + d_i)$, $i = 1, 2, \ldots$. Moreover, it spends an $\mathsf{Exp}(b_0)$ amount of time in state 0 and an $\mathsf{Exp}(b_i + d_i)$ amount of time in state $i \neq 0$.

**Theorem A.11.2 (Kolmogorov Equations)** *Any transition function $P_t$ with conservative $Q$-matrix $Q$ satisfies the* **Kolmogorov backward equations***:*

$$P_t' = Q\,P_t, \quad t \geqslant 0\,. \tag{A.51}$$

This is easy to see when $P_t$ and $Q$ are finite-dimensional, as, by the Chapman–Kolmogorov equations (A.37), $\lim_{h\downarrow 0}(P_{t+h} - P_t)/h = \lim_{h\downarrow 0}(P_h - I)/h\,P_t = Q\,P_t$. In a similar way, finite-dimensional transition functions satisfy the **Kolmogorov forward equations**:

$$P_t' = P_t\,Q, \quad t \geqslant 0\,. \tag{A.52}$$

The proof for infinite-dimensional transition functions is not as straightforward and requires certain regularity conditions on $Q$ — for example, $Q$ being conservative, as in Theorem A.11.2. Indeed, for some transition functions the forward equations may not hold at all. However, a converse result to the above theorem is as follows [1, Page 70].

**Theorem A.11.3 (Minimal Transition Function)** *For any stable $Q$-matrix $Q$ there exists a transition function $P_t^M$ that is the solution to both the backward and forward equations and is minimal in the sense that $P_t^M \leqslant P_t$ for any other solution $P_t$ of either the backward or forward equation. If $P_t^M$ is honest, it is the unique solution to the backward and forward equations.*

The Markov jump process with $P_t^M$ as its transition function is called the **minimal $Q$-process** and corresponds to the Markov jump process $X$ in Theorem A.11.1.

For a Markov jump process we usually only have knowledge of the $Q$-matrix $Q$, and so directly verifying whether or not $P_t^M$ is honest may not be easy or even possible. However, it is often possible to determine the honesty of $P_t^M$ indirectly via inspection of $Q$, as is seen from the following theorem [1].

**Theorem A.11.4 (Regular $Q$-matrix)** *If $Q$ is regular then the minimal solution to the Kolmogorov backward equations is honest, and is therefore the unique solution to the forward and backward equation. In particular, this is the case when $Q$ is conservative and uniformly bounded.*

In most applications the Markov jump process is defined by a conservative uniformly bounded $Q$-matrix (in particular, when the $Q$-matrix is of finite dimensions). The transition matrix (function) is then the unique solution to the Kolmogorov differential equations, and can be written in matrix-exponential form as

$$P_t = \mathrm{e}^{Qt} = \sum_{k=0}^{\infty} \frac{t^k Q^k}{k!} \ .$$

### A.11.1 Limiting Behavior

The limiting behavior of Markov jump processes is akin to that of the Markov chains discussed in Section A.10.2.

**Theorem A.11.5 (Limiting Distribution)** *Let $\{X_t, t \geqslant 0\}$ be an irreducible Markov jump process with regular $Q$-matrix $Q$. Then, irrespective of $x$,*

$$\lim_{t \to \infty} \mathbb{P}(X_t = y \,|\, X_0 = x) = \pi(y) \ , \tag{A.53}$$

*for some number $\pi(y) \geqslant 0$. Moreover, the row vector $\boldsymbol{\pi} = \{\pi(y)\}$ is the solution to*

$$\boldsymbol{\pi} Q = \mathbf{0}, \quad \sum_{y \in E} \pi(y) = 1 \ , \tag{A.54}$$

*provided such a solution exists, in which case all states are positive recurrent. If such a solution does not exist, then $\boldsymbol{\pi} = \mathbf{0}$.*

As in the Markov chain case, $\boldsymbol{\pi}$ defines the **limiting distribution** of $X$. Any solution $\boldsymbol{\pi}$ of (A.54) with $\sum_y \pi(y) = 1$ is called a **stationary distribution**, because taking it as the initial distribution of the Markov jump process renders the process stationary. Equations (A.54) are, as in the Markov chain case, called the **global balance equations**, and can be written as

$$\sum_{y \neq x} \pi(x)\, q(x,y) = \sum_{y \neq x} \pi(y)\, q(y,x) \quad \text{for all } x \in E \ , \tag{A.55}$$

balancing the "probability flux" out of $x$ with that into $x$. The global balance equations are readily generalized to (A.47), replacing the transition probabilities with transition rates. More importantly, if the process is *reversible* then the stationary distribution can be found from the **detailed balance equations**:

$$\pi(x)\, q(x,y) = \pi(y)\, q(y,x)\ , \quad x, y \in E\ . \tag{A.56}$$

Reversibility can be easily verified by checking that "looping" does not occur, that is, via Kolmogorov's criterion (A.49), replacing the probabilities $p$ with rates $q$. The criterion in this case is thus given by

$$q(x_1,x_2)\, q(x_2,x_3) \cdots q(x_{n-1},x_n)\, q(x_n,x_1) = q(x_1,x_n)\, q(x_n,x_{n-1}) \cdots q(x_2,x_1)$$

for all finite loops of states $x_1, \ldots, x_n, x_1$.

■ **EXAMPLE A.11** **($M/M/1$ Queue Continued)**

Let $X_t$ denote the number of customers in an $M/M/1$ queueing system at time $t \geqslant 0$; see Examples A.8 and A.9. The process $\{X_t, t \geqslant 0\}$ is an irreducible birth and death process with birth rates $\lambda$ and death rates $\mu$. The system of equations (A.54) has a unique solution

$$\pi(y) = (1-\varrho)\varrho^y, \quad y = 0, 1, 2, \ldots\ , \tag{A.57}$$

where $\varrho = \lambda/\mu$, if and only if $\varrho < 1$. For $\lambda < \mu$ all the states are therefore positive recurrent. Note that any birth and death process is reversible. As a consequence (A.57) can be found directly from the local balance equations

$$\pi(y)\, \lambda = \pi(y+1)\, \mu, \quad y = 0, 1, \ldots\ .$$

Theorem A.9.1 shows that for $\varrho < 1$ the steady-state expected number of customers in the system is $\mathbb{E}X = \varrho/(1-\varrho)$.

## A.12 ITÔ INTEGRAL AND ITÔ PROCESSES

An important class of stochastic processes — that of **Itô processes** — is constructed from the Wiener process via the notion of the Itô integral. The Wiener process is discussed in more detail in Section 5.5, but here we only consider its role in Itô integration. The Itô integral provides the mathematical justification of integrals of the form ☞ 177

$$\int_0^T F_t\, \mathrm{d}W_t\ ,$$

where $W = \{W_t\}$ is a Wiener process and $F = \{F_t\}$ is a stochastic process. In its simplest form the Itô integral is defined for processes $F$ that are **predictable** [18] with respect to the history of $W$ and satisfy

$$\mathbb{E}\int_0^T F_s^2\, \mathrm{d}s < \infty\ . \tag{A.58}$$

We will denote this class of integrands by $\mathscr{H}_T$. A sufficient condition for predictability is that the process is left-continuous and adapted — so, $F_t$ may depend on $\{W_s, s \leqslant t\}$ but not on $\{W_s, s > t\}$. Let $t \leqslant T$ and $F \in \mathscr{H}_T$. The **Itô integral** of $F$ with respect to $W$ over $[0, t]$ is defined as

$$\int_0^t F_s \, \mathrm{d}W_s \overset{\text{def}}{=} \lim_{n \to \infty} \sum_{k=0}^{n-1} F_{t_k}(W_{t_{k+1}} - W_{t_k}), \quad 0 = t_0 < \cdots < t_n = t \, , \tag{A.59}$$

where $\lim_{n\to\infty} \max_k\{t_{k+1} - t_k\} = 0$, and the convergence is in the mean square sense (see Section A.8.1).

**Remark A.12.1 (Stochastic Integral)** The Itô integral is an example of a **stochastic integral**. The general theory of stochastic integration [21, 23] allows $\{W_t\}$ to be replaced by **semimartingales** — processes that can be decomposed as the sum of a (local) martingale and a process of finite variation — and the integrand process $\{F_t\}$ by predictable processes that satisfy weaker conditions than (A.58). In particular, it can be shown that the limit (A.59) still exists, but in probability rather than in the mean square sense, if (A.58) is replaced by

$$\int_0^T F_s^2 \, \mathrm{d}s < \infty \quad \text{a.s.} \tag{A.60}$$

An **Itô process** is any stochastic process $\{X_t, 0 \leqslant t \leqslant T\}$ that can be written in the form

$$X_t = X_0 + \int_0^t \mu_s \, \mathrm{d}s + \int_0^t \sigma_s \, \mathrm{d}W_s, \quad 0 \leqslant t \leqslant T \, ,$$

where $\{\mu_t\}$ is adapted, with $\int_0^T |\mu_t| \, \mathrm{d}t < \infty$ and $\{\sigma_t\} \in \mathscr{H}_T$. The above integral equation is usually written in the shorthand differential form

$$\mathrm{d}X_t = \mu_t \, \mathrm{d}t + \sigma_t \, \mathrm{d}W_t \, . \tag{A.61}$$

Note that the coefficients $\mu_t$ and $\sigma_t$ may depend on the whole path $\{W_s, s \leqslant t\}$. An **$m$-dimensional Itô process** $\{\mathbf{X}_t\} = \{(X_{t,1}, \ldots, X_{t,m})^\top\}$ driven by an $n$-dimensional Wiener process $\{\mathbf{W}_t\} = \{(W_{t,1}, \ldots, W_{t,n})^\top\}$ can be defined analogously via the differential expression

$$\mathrm{d}X_{t,i} = \mu_{t,i} \, \mathrm{d}t + \sum_{j=1}^{n} \sigma_{t,ij} \, \mathrm{d}W_{t,j}, \quad i = 1, \ldots, m \, ,$$

written in matrix–vector notation as

$$\mathrm{d}\mathbf{X}_t = \boldsymbol{\mu}_t \, \mathrm{d}t + \boldsymbol{\sigma}_t \, \mathrm{d}\mathbf{W}_t \, , \tag{A.62}$$

where

$$\boldsymbol{\mu}_t = \begin{pmatrix} \mu_{t,1} \\ \vdots \\ \mu_{t,m} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\sigma}_t = \begin{pmatrix} \sigma_{t,11} & \cdots & \sigma_{t,1n} \\ \vdots & \ddots & \vdots \\ \sigma_{t,m1} & \cdots & \sigma_{t,mn} \end{pmatrix} .$$

An Itô process is an example of a semimartingale. As a special case of the general theory of stochastic integration with respect to such processes one may define integration with respect to Itô processes. In particular, (see, for example, [18]) if $X = \{X_t\}$ is an Itô process and $F = \{F_t\} \in \mathscr{H}_T$, then the stochastic integral of $F$ with respect to $X$ is defined as:

$$\int_0^t F_s \mathrm{d}X_s \overset{\mathrm{def}}{=} \int_0^t F_s\, \mu_s\, \mathrm{d}s + \int_0^t F_s\, \sigma_s\, \mathrm{d}W_s, \quad 0 \leqslant t \leqslant T\,.$$

Let $X = \{X_t\}$ and $Y = \{Y_t\}$ be two processes adapted to the same filtration. Then,

$$[X, Y]_t \overset{\mathrm{def}}{=} \lim_{n\to\infty} \sum_{k=0}^{n-1} (X_{t_{k+1}} - X_{t_k})(Y_{t_{k+1}} - Y_{t_k})\,,$$

where $0 = t_0 < \cdots < t_n = t$ and $\lim_{n\to\infty} \max_k\{t_{k+1} - t_k\} = 0$, is called the **covariation** between the processes $X$ and $Y$. The special case $[X, X]_t$, denoted $[X]_t$, is called the **quadratic variation** of $X$.

Below we list a number of properties of Itô integrals and Itô processes. Proofs may be found in [23], for example.

1. *Isometry property*: If $F, G \in \mathscr{H}_T$, then for any $0 \leqslant t \leqslant T$,

$$\mathbb{E} \int_0^t F_s\, \mathrm{d}W_s \int_0^t G_s\, \mathrm{d}W_s = \mathbb{E} \int_0^t F_s\, G_s\, \mathrm{d}s\,.$$

2. *Martingale property*: If $F \in \mathscr{H}_T$, then the Itô process defined by

$$Y_t = \int_0^t F_s\, \mathrm{d}W_s, \quad 0 \leqslant t \leqslant T\,,$$

is a square-integrable martingale.

3. *Quadratic variation and covariation*: Let $\mathrm{d}X_t = \mu_t\, \mathrm{d}t + \sigma_t\, \mathrm{d}W_t$ and $\mathrm{d}Y_t = \nu_t\, \mathrm{d}t + \varrho_t\, \mathrm{d}W_t$ define two Itô processes with respect to the *same* Wiener process $\{W_t\}$. Then,

$$[X, Y]_t = \int_0^t \sigma_s\, \varrho_s\, \mathrm{d}s\,.$$

In shorthand differential form the covariation and the quadratic variation are

$$\mathrm{d}[X, Y]_t = \sigma_t\, \varrho_t\, \mathrm{d}t \quad \text{and} \quad \mathrm{d}[X]_t = \sigma_t^2\, \mathrm{d}t\,, \quad \text{respectively.}$$

4. *Covariance for multivariate Itô process*: Let $\{\mathbf{X}_t\}$ be an $m$-dimensional Itô process. Then using the formal rules (see [18]) $(\mathrm{d}t)^2 = \mathrm{d}t\, \mathrm{d}W_{t,i} = 0$ and $\mathrm{d}W_{t,i}\, \mathrm{d}W_{t,j} = \delta_{ij}\, \mathrm{d}t$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, we can write:

$$\mathrm{d}[X_{\cdot,i}, X_{\cdot,j}]_t = \mathrm{d}X_{t,i}\, \mathrm{d}X_{t,j} = \sum_{k=1}^{n} \sigma_{t,ik}\, \sigma_{t,jk}\, \mathrm{d}t\,, \quad i, j \in \{1, \ldots, m\}\,.$$

5. *Itô's lemma*: Let $\mathrm{d}X_t = \mu_t\, \mathrm{d}t + \sigma_t\, \mathrm{d}W_t$ define an Itô process and let $f(x) : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable with first and second derivatives $f'$ and $f''$, respectively. Then,

$$f(X_t) = f(X_0) + \int_0^t f'(X_s)\, \mathrm{d}X_s + \frac{1}{2} \int_0^t f''(X_s)\, \sigma_s^2\, \mathrm{d}s \tag{A.63}$$

or, in differential form:

$$\mathrm{d}f(X_t) = f'(X_t)\,\mathrm{d}X_t + \frac{1}{2}f''(X_t)\,\sigma_t^2\,\mathrm{d}t\,.$$

Compare this with the corresponding **chain rule** of ordinary calculus: $\mathrm{d}f(x(t)) = f'(x(t))\,\mathrm{d}x(t)$.

6. *Itô's lemma in* $\mathbb{R}^m$: Let $\{\mathbf{X}_t\}$ be an $m$-dimensional Itô process, and $f : \mathbb{R}^m \to \mathbb{R}$ be twice continuously differentiable in all variables, then

$$\mathrm{d}f(\mathbf{X}_t) = \sum_{i=1}^{m} \partial_i f(\mathbf{X}_t)\,\mathrm{d}X_{t,i} + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \partial_{ij} f(\mathbf{X}_t)\,\mathrm{d}[X_{\cdot,i}, X_{\cdot,j}]_t\,. \tag{A.64}$$

A special case is the **product rule** for Itô processes:

$$\mathrm{d}(X_t Y_t) = Y_t\,\mathrm{d}X_t + X_t\,\mathrm{d}Y_t + \mathrm{d}[X,Y]_t\,. \tag{A.65}$$

The corresponding integral form is the Itô **integration by parts** formula.

Another special case is $\mathbf{X}_t = (X_t, t)^\top$, where $t$ ($\geqslant 0$) is deterministic and the process $\{X_t\}$ is governed by $\mathrm{d}X_t = \mu_t\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t$. Then,

$$\mathrm{d}f(X_t, t) = \left(\frac{\partial f}{\partial t}(\mathbf{X}_t) + \mu_t \frac{\partial f}{\partial x}(\mathbf{X}_t) + \frac{\sigma_t^2}{2}\frac{\partial^2 f}{\partial x^2}(\mathbf{X}_t)\right)\mathrm{d}t + \sigma_t \frac{\partial f}{\partial x}(\mathbf{X}_t)\,\mathrm{d}W_t\,. \tag{A.66}$$

7. *Gaussian process for deterministic integrands*: If $f(t,s)$ is a nonrandom function with $\int_0^t f^2(t,s)\,\mathrm{d}s < \infty$ for any $0 \leqslant t \leqslant T$, then the Itô integral

$$Y_t = \int_0^t f(t,s)\,\mathrm{d}W_s \tag{A.67}$$

defines a Gaussian process $\{Y_t, 0 \leqslant t \leqslant T\}$ with mean zero and covariance function

$$\mathrm{Cov}(Y_s, Y_t) = \int_0^{\min\{s,t\}} f(t,u) f(s,u)\,\mathrm{d}u\,.$$

Note that, unless $f(t,s) = f(s)$, $\{Y_t\}$ need not be a martingale. If $f(t,s) = f(s)$, then by the integration by parts formula we also have

$$Y_t = W_t f(t) - \int_0^t W_s\,\mathrm{d}f(s)\,.$$

8. *Time-change*: Let $\{W_t\}$ be a Wiener process, and define $Z_t = W_{C_t}$, $t \geqslant 0$, for some given deterministic function $C_t = \int_0^t f^2(s)\,\mathrm{d}s < \infty$ for all $t \leqslant T$. Then, the stochastic process $\{Z_t, 0 \leqslant t \leqslant T\}$ has the same distribution as the Itô integral process $\{Y_t\}$ defined in (A.67) with $f(t,s) = f(s)$.

9. *Girsanov's theorem*: Let $\mathrm{d}\mathbf{X}_t = \boldsymbol{\mu}_t\,\mathrm{d}t + \mathrm{d}\mathbf{W}_t$ define a multidimensional Itô process under probability measure $\mathbb{P}$ with respect to a filtration $\mathcal{F} =$

$\{\mathcal{F}_t, t \geqslant 0\}$. Assume that $\{\boldsymbol{\mu}_t, t \geqslant 0\}$ satisfies **Novikov's condition**: $\mathbb{E}\exp(\frac{1}{2}\int_0^t \boldsymbol{\mu}_s^\top \boldsymbol{\mu}_s \, \mathrm{d}s) < \infty$. For each $t \geqslant 0$ define

$$M_t = \exp\left(\int_0^t \boldsymbol{\mu}_s^\top \, \mathrm{d}\mathbf{W}_s - \frac{1}{2}\int_0^t \boldsymbol{\mu}_s^\top \boldsymbol{\mu}_s \, \mathrm{d}s\right) .$$

Then $\{M_t, t \geqslant 0\}$ is a martingale with respect to $\mathcal{F}$. For a fixed $T \geqslant 0$ let $\mathbb{P}_T$ denote the restriction of $\mathbb{P}$ to $\mathcal{F}_T$. Define a new measure $\widetilde{\mathbb{P}}_T$ by

$$\widetilde{\mathbb{P}}_T(A) = \mathbb{E}_T M_T \mathrm{I}_A \,, \quad A \in \mathcal{F}_T \,,$$

so that

$$\mathbb{P}_T(A) = \widetilde{\mathbb{E}}_T \mathrm{I}_A / M_T \,.$$

Then under $\widetilde{\mathbb{P}}_T$ the process $\{\mathbf{X}_t, 0 \leqslant t \leqslant T\}$ is a Wiener process.

**Remark A.12.2 (Stratonovich Integral)** Let $W$ be a Wiener process and $X \in \mathscr{H}_T$. For any $0 \leqslant t \leqslant T$ let

$$\int_0^t X_s \circ \mathrm{d}W_s \overset{\text{def}}{=} \lim_{n\to\infty} \sum_{k=0}^{n-1} \frac{X_{t_k} + X_{t_{k+1}}}{2}(W_{t_{k+1}} - W_{t_k}), \quad 0 = t_0 < \cdots < t_n = t \,,$$

where $\lim_{n\to\infty} \max_k\{t_{k+1} - t_k\} = 0$ and convergence is in the mean square sense. This defines the **Stratonovich integral** of $X$ with respect to $W$ over $[0, t]$. This integral does not in general define a martingale, and therefore most of the properties above do not directly apply. However, the Stratonovich integral has the advantage that it formally obeys the standard calculus formulas. In particular, for a three times continuously differentiable function $f$, the Stratonovich integral formally satisfies the ordinary chain rule

$$\mathrm{d}f(X_t) = f'(X_t) \circ \mathrm{d}X_t \,.$$

## A.13 DIFFUSION PROCESSES

Let $\{W_t\}$ be a Wiener process, and $a(x,t)$ and $b(x,t)$ be deterministic functions. A **stochastic differential equation** (SDE) for a stochastic process $\{X_t\}$ is an expression of the form

$$\mathrm{d}X_t = a(X_t, t)\,\mathrm{d}t + b(X_t, t)\,\mathrm{d}W_t \,. \tag{A.68}$$

The coefficient $a$ is called the **drift** and $b^2$ (or sometimes $b$) the **diffusion** coefficient. When $a$ and $b$ do not depend on $t$ explicitly (that is, $a(x,t) = \widetilde{a}(x)$, and $b(x,t) = \widetilde{b}(x)$), the SDE is said to be **autonomous** or **homogeneous**. When $a$ and $b$ are linear in $x$, the SDE is said to be **linear**.

Intuitively, the process $\{X_t\}$ is specified by a "noisy ODE", relating its derivative at $t$ to a function of its present value $X_t$ and an additional noise term. Mathematically, $\{X_t\}$ is the solution to the integral equation

$$X_t = X_0 + \int_0^t a(X_s, s)\,\mathrm{d}s + \int_0^t b(X_s, s)\,\mathrm{d}W_s \,, \tag{A.69}$$

where the last integral is defined in the Itô sense. Note that when $b \equiv 0$, we obtain an ordinary differential equation.

**Remark A.13.1 (Diffusion-Type SDE)** Although SDEs of the type above are by far the most common, it should be noted that there exist more general SDEs [18, 19], where, for example, $a$ and $b$ depend on the whole history of $\{X_s, s \leqslant t\}$ rather than only on $t$ and $X_t$. The special case (A.68) is also referred to as a **diffusion-type** SDE.

A stochastic process $\{X_t\}$ is said to be a **strong solution** to the SDE (A.68) if $X_t$ is a function of $t$ and the underlying Wiener process $\{W_s, s \leqslant t\}$, and satisfies (A.69). It is called a **weak solution** if (A.69) holds for *some* Wiener process.

The following theorem gives conditions for existence and uniqueness of strong solutions on an interval $[0, T]$. A proof can be found, for example, in [19].

**Theorem A.13.1 (Existence and Uniqueness of Strong Solutions)** *Suppose the following conditions are satisfied:*

1. **Linear growth condition**: *There is a constant $C$ such that for all $t \in [0,T]$*

$$|a(x,t)| + |b(x,t)| \leqslant C(1 + |x|) \quad \text{for all } x \, . \tag{A.70}$$

2. **Local Lipschitz continuity in $x$**: *For every $K > 0$ there is a constant $D_K$ such that for all $t \in [0,T]$*

$$|a(x,t) - a(y,t)| + |b(x,t) - b(y,t)| \leqslant D_K|x - y| \quad \text{for all } x, y \in [-K, K] \, . \tag{A.71}$$

3. *$X_0$ is independent of $\{W_t, 0 \leqslant t \leqslant T\}$ and has finite variance.*

*Then the SDE (A.68) has a unique strong solution on $[0,T]$. In addition, the solution has almost surely continuous paths, is a strong Markov process, and $\int_0^T \mathbb{E} X_s^2 \, \mathrm{d}s < \infty$.*

The linear growth condition ensures that each path of the SDE does not "explode"; that is, the path does not tend to $\pm\infty$ within a finite interval of time. Note that a similar condition is required for ordinary differential equations. For example, the differential equation $\mathrm{d}x(t) = x^2(t)\,\mathrm{d}t$, $x(0) = a$ has a "local" solution $x(t) = a/(1 - at)$ on the interval $[0, 1/a)$ rather than a "global" solution on $\mathbb{R}_+$. Removing Condition 1 still gives a unique strong solution, but only up to a (random) time of explosion.

Local Lipschitz continuity ensures that solutions of SDEs can be constructed via an iterative procedure, similar to that for ordinary differential equations (Picard iteration [23]). As a measure of smoothness of a function, this condition lies between continuity and differentiability. In particular, if $a$ and $b$ are continuously differentiable in $x$, or, more generally, if their derivatives in $x$ are uniformly bounded on $[0, T]$, then they satisfy (A.71).

Weak solutions to the SDE (A.68) exist under slightly more general conditions; see for example [23]. In particular, if for each $t$ the functions $a$ and $b$ are bounded and continuous in $x$. Such solutions are only defined through their probability distributions, rather than pathwise via the Wiener process $W$.

■ **EXAMPLE A.12 (Linear SDE)**

For a linear SDE

$$\mathrm{d}X_t = (\alpha_t + \beta_t X_t)\,\mathrm{d}t + (\gamma_t + \delta_t X_t)\,\mathrm{d}W_t \ ,$$

the (strong) solution can be given explicitly as the product $X_t = U_t\, V_t$, with

$$U_t = \exp\left\{\int_0^t \left(\beta_s - \frac{1}{2}\delta_s^2\right)\mathrm{d}s + \int_0^t \delta_s\,\mathrm{d}W_s\right\} ,$$

$$V_t = X_0 + \int_0^t \frac{\alpha_s - \gamma_s\delta_s}{U_s}\mathrm{d}s + \int_0^t \frac{\gamma_s}{U_s}\,\mathrm{d}W_s \ .$$

In particular, if $\delta_t \equiv 0$ and $\beta_t \equiv \beta$ (constant), then

$$X_t = \mathrm{e}^{\beta t}\left(X_0 + \int_0^t \mathrm{e}^{-\beta s}\alpha_s\,\mathrm{d}s + \int_0^t \mathrm{e}^{-\beta s}\gamma_s\,\mathrm{d}W_s\right) ,$$

and $\{X_t\}$ is therefore a Gaussian process, provided that the distribution of $X_0$ is Gaussian (this includes the case where $X_0$ is a constant). See [19, Pages 110–113] for more details.

A solution $\{X_t\}$ to (A.68) or, more precisely to (A.69), is called a **diffusion process**, or, more specifically, an **Itô diffusion**. From Theorem A.13.1, Itô diffusions are Markov processes with continuous paths. Let $X = \{X_t\}$ be an Itô diffusion with drift and diffusion coefficients $a$ and $b^2$, respectively. The meaning of these terms becomes clear when considering the infinitesimal behavior of $X$. In particular, by (A.69),

$$X_{t+h} - X_t = \int_t^{t+h} a(X_s, s)\,\mathrm{d}s + \int_t^{t+h} b(X_s, s)\,\mathrm{d}W_s \ .$$

Taking the conditional expectation given $X_t = x$ on both sides yields

$$\mathbb{E}[X_{t+h} - x \,|\, X_t = x] = a(x,t)\,h + o(h) \ ,$$

since the expectation of the second integral in the above integral equation is 0, due to the martingale property of the Itô integral. Similarly, using the isometry Property 1 on Page 641,

$$\mathrm{Var}(X_{t+h} - x \,|\, X_t = x) = \mathbb{E}[(X_{t+h} - x - a(x,t)\,h)^2 \,|\, X_t = x] + o(h) = b^2(x,t)\,h + o(h) \ .$$

In other words, given that the process is at position $x$ at time $t$, the displacement of $X$ in the next $h \ll 1$ time units has expectation $a(x,t)h$ and variance $b^2(x,t)h$.

**Remark A.13.2 (Boundary Behavior)** We have considered only diffusions on the whole real line. Diffusions on a half-line or intervals are also possible. For such processes the behavior at the boundary needs to be specified, in addition to the behavior in the interior of the domain described by the SDE. See, for example [8, 16].

The analogue of (A.68) in $\mathbb{R}^m$ is given by the **multidimensional SDE**

$$\mathrm{d}\mathbf{X}_t = \mathbf{a}(\mathbf{X}_t, t)\,\mathrm{d}t + B(\mathbf{X}_t, t)\,\mathrm{d}\mathbf{W}_t \ , \tag{A.72}$$

where $\{\mathbf{W}_t\}$ is an $n$-dimensional Wiener process, $\mathbf{a}(\mathbf{x}, t)$ is an $m$-dimensional vector (the drift) and $B(\mathbf{x}, t)$ an $m \times n$ matrix, for each $\mathbf{x} \in \mathbb{R}^m$ and $t \in \mathbb{R}$. The $m \times m$ matrix $C = BB^\top$ is called the **diffusion matrix**.

As with the one-dimensional case, existence and uniqueness of strong solutions to multidimensional SDEs relies on certain Lipschitz and linear growth conditions. In particular, we have the following multidimensional version of Theorem A.13.1 (see [18, Page 173]):

**Theorem A.13.2 (Strong Solutions of Multidimensional SDEs)** *Suppose the following conditions are satisfied, where for a matrix $A$, $\|A\| \overset{\text{def}}{=} \sqrt{\operatorname{tr}(AA^\top)}$:*

*1.* **Linear growth condition***: There is a constant $C$ such that for all $t \in [0, T]$*

$$\|\mathbf{a}(\mathbf{x}, t)\| + \|B(\mathbf{x}, t)\| \leqslant C(1 + \|\mathbf{x}\|) \quad \textit{for all } \mathbf{x} \,. \tag{A.73}$$

*2.* **Local Lipschitz continuity in x***: For every $K > 0$ there is a constant $D_K$ such that for all $t \in [0, T]$*

$$\|\mathbf{a}(\mathbf{x}, t) - \mathbf{a}(\mathbf{y}, t)\| + \|B(\mathbf{x}, t) - B(\mathbf{y}, t)\| \leqslant D_K \|\mathbf{x} - \mathbf{y}\| \quad \textit{for all } \|\mathbf{x}\|, \|\mathbf{y}\| \leqslant K \,. \tag{A.74}$$

*3. $\mathbf{X}_0$ is independent of $\{\mathbf{W}_t, 0 \leqslant t \leqslant T\}$ and $\mathbb{E}\|\mathbf{X}_0\|^2 < \infty$.*

*Then the SDE* (A.68) *has a unique strong solution on* $[0, T]$.

### A.13.1 Kolmogorov Equations

For autonomous SDEs, that is, those of the form

$$\mathrm{d}X_t = a(X_t)\,\mathrm{d}t + b(X_t)\,\mathrm{d}W_t \,, \tag{A.75}$$

the corresponding diffusion process is a *time-homogeneous* Markov process. The corresponding transition kernel $P_t$ can be found from the Kolmogorov backward (and under more restrictive conditions) from the Kolmogorov forward equations. To see this, let $L$ be the linear elliptic differential operator

$$Lf(x) = a(x)f'(x) + \frac{1}{2}b^2(x)f''(x) \tag{A.76}$$

acting on all twice continuously differentiable functions on compact sets. Then, by Itô's formula,

$$M_t \overset{\text{def}}{=} f(X_t) - f(X_0) - \int_0^t Lf(X_s)\,\mathrm{d}s$$

defines a martingale on $[0, T]$. In particular, denoting by $\mathbb{E}^x$ the expectation operator under which the process starts at $x$, we have

$$\mathbb{E}^x f(X_t) = f(x) + \int_0^t \mathbb{E}^x Lf(X_s)\,\mathrm{d}s \,, \tag{A.77}$$

where the interchange of expectation and integral is allowed by Fubini's theorem. It follows that

$$Lf(x) = \lim_{t \downarrow 0} \frac{\mathbb{E}^x f(X_t) - f(x)}{t} \,. \tag{A.78}$$

The limit in (A.78) also defines the **infinitesimal generator** of the Markov process. The domain of the infinitesimal generator consists of all bounded measurable functions for which the limit exists — this includes the domain of $L$, hence the infinitesimal generator extends $L$. Let $P_t$ be the transition kernel of the Markov process and define the operator $P_t$ by

$$P_t f(x) = \int P_t(x, \mathrm{d}y) f(y) = \mathbb{E}^x f(X_t) \,.$$

Then, by (A.77), we obtain the **Kolmogorov forward equations**:

$$P_t' f = P_t L f \,. \tag{A.79}$$

Moreover, by the Chapman–Kolmogorov equations we have $P_{t+s} f(x) = P_s P_t f(x) = \mathbb{E}^x P_t f(X_s)$, and therefore

$$\frac{1}{s}\left\{P_{t+s} f(x) - P_t f(x)\right\} = \frac{1}{s}\left\{\mathbb{E}^x P_t f(X_s) - P_t f(x)\right\} \,.$$

Letting $s \downarrow 0$, we obtain the **Kolmogorov backward equations**:

$$P_t' f = L P_t f \,. \tag{A.80}$$

If $P_t$ has a transition density $p_t$, then we can write the last equation as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int p_t(x,y) f(y) \,\mathrm{d}y = \int L p_t(x,y) f(y) \,\mathrm{d}y \,,$$

so that $p_t(x, y)$ for fixed $y$ satisfies the Kolmogorov backward equations:

$$\begin{aligned} \frac{\partial}{\partial t} p_t(x,y) &= L p_t(x,y) \\ &= a(x) \frac{\partial}{\partial x} p_t(x,y) + \frac{1}{2}\, b^2(x) \frac{\partial^2}{\partial x^2}\, p_t(x,y) \,. \end{aligned} \tag{A.81}$$

Similarly, (A.79) can be written as $\frac{\mathrm{d}}{\mathrm{d}t} \int p_t(x,y) f(y) \,\mathrm{d}y = \int p_t(x,y) L f(y) \,\mathrm{d}y = \int f(y) L^* p_t(x,y) \,\mathrm{d}y$, where $L^*$ (acting here on $y$) is the adjoint operator of $L$ defined by $\int g(y) L h(y) \,\mathrm{d}y = \int h(y) L^* g(y) \,\mathrm{d}y$. Hence, for fixed $x$ the density $p_t(x, y)$ satisfies the Kolmogorov forward equations, also called the **Fokker–Planck equations**:

$$\begin{aligned} \frac{\partial}{\partial t} p_t(x,y) &= L^* p_t(x,y) \\ &= -\frac{\partial}{\partial y} \left(a(y)\, p_t(x,y)\right) + \frac{1}{2} \frac{\partial^2}{\partial y^2} \left(b^2(y)\, p_t(x,y)\right) \,. \end{aligned} \tag{A.82}$$

Sufficient conditions on $a(x)$ and $b(x)$ such that $p_t(x, y)$ exists and is the unique solution to the forward and backward equations are that $a(x)$ and $b(x)$ have partial derivatives up to order two, which are bounded and satisfy a Lipschitz condition; see also [18]. This illustrates the important connection between partial differential equations of the form $u_t' = L u_t$ and diffusion processes. Indeed, given an elliptic operator $L$, the pdf of the corresponding diffusion process gives the fundamental solution (Green's function) of the partial differential equation — see also Chapter 17.

☞ 577

**Remark A.13.3 (Operators for Multidimensional SDEs)** For multidimensional SDEs of the form (A.72) the infinitesimal generator extends the operator

$$Lf(\mathbf{x}) = \sum_{i=1}^{m} a_i(\mathbf{x}) \frac{\partial}{\partial x_i} f(\mathbf{x}) + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} C_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \, \partial x_j} f(\mathbf{x}) \, ,$$

where $\{a_i\}$ are the components of $\mathbf{a}$ and $\{C_{ij}\}$ the components of $C = BB^\top$.

## A.13.2 Stationary Distribution

Consider again the diffusion governed by the autonomous SDE (A.75). Suppose that

$$\pi(y) = \int p_t(x, y) \, \pi(x) \, \mathrm{d}x \, ,$$

where $p_t$ is the transition density of the diffusion. Then $\pi(x)$ is called a **stationary** or **invariant density** of the diffusion (A.75). If the initial state $X_0$ has density $\pi(x)$, then $\{X_t, t \geqslant 0\}$ is a stationary process.

**Theorem A.13.3 (Stationary Distribution)** *If the stationary density of* (A.75) *exists and is twice continuously differentiable, then it solves the ODE*

$$L^* \pi = 0 \quad \Leftrightarrow \quad \frac{1}{2} \frac{\mathrm{d}^2}{\mathrm{d}y^2} \left( b^2(y) \, \pi(y) \right) - \frac{\mathrm{d}}{\mathrm{d}y} \left( a(y) \, \pi(y) \right) = 0 \, ,$$

*where $L^*$ is the adjoint operator in* (A.82). *The stationary density that solves the ODE is of the form*

$$\pi(x) = \frac{c}{b^2(x)} \exp\left( \int_{x_0}^{x} \frac{2\, a(y)}{b^2(y)} \mathrm{d}y \right) ,$$

*where $x_0$ is an arbitrary constant and $c$ is a constant such that $\int \pi(y) \, \mathrm{d}y = 1$.*

For a rigorous discussion of the conditions for existence of $\pi$, see [22].

Loosely speaking, if the diffusion process is in the stationary regime, its distribution does not change in time. Hence, the transition density $p_t$ is independent of time and the partial derivative with respect to $t$ in (A.82) is zero, giving the equation $L^* \pi = 0$ satisfied by the stationary density.

## A.13.3 Feynman–Kac Formula

The Feynman–Kac formula establishes an important relationship between stochastic processes and linear parabolic PDEs. The result can be used to approximate the solution of a PDE via Monte Carlo methods. Alternatively, conditional expectations of a diffusion process can be computed by solving a PDE, see Chapter 17.

☞ 577

For each $t \geqslant 0$ let $L_t$ be the linear elliptic differential operator

$$L_t u(x, t) = a(x, t) \frac{\partial}{\partial x} u(x, t) + \frac{1}{2} \, b^2(x, t) \frac{\partial^2}{\partial x^2} u(x, t)$$

acting on all twice continuously differentiable functions on compact sets.

**Theorem A.13.4 (Feynman–Kac Formula)** *Let $k(x,t)$ and $f(x)$ be bounded functions and let the process $\{X_t, 0 \leqslant t \leqslant T\}$ evolve according to* (A.68). *Assume that the solution to the PDE*

$$\left(L_t + \frac{\partial}{\partial t} - k(x,t)\right) u(x,t) = 0, \quad x \in \mathbb{R},\ t \in [0,T]\ ,$$

*with final condition $u(x,T) = f(x)$ exists. Then, the solution is unique and given by*

$$u(x,t) = \mathbb{E}\left[\mathrm{e}^{-\int_t^T k(X_s,s)\,\mathrm{d}s} f(X_T)\,\middle|\, X_t = x\right], \quad t \in [0,T]\ .$$

We explain why the formula is plausible (for a detailed treatment see [10, 22]). Define the process $\{Y_t\}$ via $Y_t = \mathrm{e}^{-\int_0^t k(X_s,s)\,\mathrm{d}s}$. Then, applying the Itô formula (A.66) we have

$$\mathrm{d}(Y_t\, u(X_t,t)) = Y_t\left(\left(L_t + \frac{\partial}{\partial t} - k(X_t,t)\right)u(X_t,t)\,\mathrm{d}t + b(X_t,t)\,\frac{\partial u}{\partial x}(X_t,t)\,\mathrm{d}W_t\right)\ .$$

Since $u$ is the solution to the PDE, the drift term is 0. Since $Y_t$ is bounded by assumption, it can be shown [10] that existence and uniqueness of the solution of the PDE implies that $\int_0^T \mathbb{E}\,|Y_t\, u(X_t,t)| < \infty$. Therefore, the process

$$Y_t\, u(X_t,t) = \int_0^t Y_s\, b(X_s,s)\,\frac{\partial u}{\partial x}(X_s,s)\,\mathrm{d}W_s$$

is a martingale. Using the Markov property of the SDE (see Theorem A.13.1) and the final condition we obtain

$$Y_t\, u(X_t,t) = \mathbb{E}[\,Y_T\, u(X_T,T)\,|\, X_s,\ 0 \leqslant s \leqslant t] = \mathbb{E}[\,Y_T\, f(X_T)\,|\, X_t]\ ,$$

which after rearrangement yields the desired result. For multidimensional analogues of the Feynman–Kac formula see Chapter 17.

### A.13.4 Exit Times

Diffusion processes are often studied through their exit times from an interval. Below we assume that $\{X_t\}$ is a homogeneous diffusion process defined by the SDE (A.75) and satisfying the existence and uniqueness conditions of Theorem A.13.1.

Let $[l,r]$ (with $l < r$) be an arbitrary interval, and let $\tau_l$ and $\tau_r$ be the first times that the process hits $l$ and $r$, respectively. Let $\tau = \min\{\tau_l, \tau_r\} = \tau_l \wedge \tau_r$ be the first **exit time** from the interval $[l,r]$.

The following results may, for example, be found in [18]. Central in the proof is the fact that, by Itô's lemma, the process $\{M_t\}$ defined by

$$M_t = f(X_{t\wedge\tau}) - \int_0^{t\wedge\tau} Lf(X_u)\,\mathrm{d}u \quad \text{is a martingale.}$$

**Theorem A.13.5 (Exit Times)** *Let the diffusion coefficient $b(x)$ be a strictly positive and continuous function on $[l,r]$, and let $f$ be any twice continuously differentiable function. Then the following holds:*

1. *The function $s$ given by $s(x) = \mathbb{E}^x\tau$ satisfies the differential equation*

$$Ls = -1, \quad \textit{with} \quad s(l) = 0, \quad s(r) = 0 \ ,$$

*where operator $L$ is given in* (A.76).

2. *Any nonconstant positive solution of $Lh = 0$ is of the form*

$$h(x; x_0, y_0) = \int_{x_0}^{x} \exp\left( - \int_{y_0}^{y} \frac{2a(u)}{b^2(u)}\,\mathrm{d}u \right) \mathrm{d}y \ ,$$

*for some arbitrary constants $x_0, y_0$. These are called* **harmonic** *functions for $L$.*

3. *For any such harmonic function,*

$$\mathbb{P}^x(\tau_l < \tau_r) = \frac{h(r) - h(x)}{h(r) - h(l)} \ .$$

## Further Reading

An easy introduction to probability theory with many examples can be found in [25]. More detailed textbooks include [11] and [28]. Classical references on probability theory are [5] and [9]. A good non-measure-theoretic introduction to stochastic processes is [24]. A detailed treatment of Markov processes can be found in [7], and a handy text on Markov processes with countable state spaces is [1]. An accessible measure-theoretic introduction to probability theory, including stochastic processes, can be found in [3]. For many examples in probability theory and stochastic processes, see Feller's two volumes [8, 9]. Other good references for stochastic processes are [4, 15, 16], and the classic [6].

## REFERENCES

1. W. J. Anderson. *Continuous-Time Markov Chains: An Applications-Oriented Approach.* Springer-Verlag, New York, 1991.
2. P. Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, New York, 1968.
3. P. Billingsley. *Probability and Measure.* John Wiley & Sons, New York, third edition, 1995.
4. E. Çinlar. *Introduction to Stochastic Processes.* Prentice Hall, Englewood Cliffs, NJ, 1975.
5. K. L. Chung. *A Course in Probability Theory.* Academic Press, New York, second edition, 1974.
6. J. L. Doob. *Stochastic Processes.* John Wiley & Sons, New York, 1953.
7. S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence.* John Wiley & Sons, New York, 1986.

8. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, New York, 1966.

9. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, New York, second edition, 1970.

10. D. Freedman. *Brownian Motion and Diffusion.* Springer-Verlag, New York, 1971.

11. G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Oxford University Press, Oxford, third edition, 2001.

12. P. R. Halmos. *Measure Theory.* Springer-Verlag, New York, second edition, 1978.

13. C. C. Heyde. On a property of the lognormal distribution. *Journal of the Royal Statistical Society, Series B*, 25(2):392–393, 1963.

14. P. L. Hsu and H. Robbins. Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences, U.S.A.*, 33(2):25–31, 1947.

15. S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, New York, second edition, 1975.

16. S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes.* Academic Press, New York, 1981.

17. F. P. Kelly. *Reversibility and Stochastic Networks.* John Wiley & Sons, New York, 1979.

18. F. C. Klebaner. *Introduction to Stochastic Calculus with Applications.* Imperial College Press, London, second edition, 2005.

19. P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations.* Springer-Verlag, Berlin, 1999. Corrected third printing.

20. J. R. Norris. *Markov Chains.* Cambridge University Press, Cambridge, 1997.

21. B. Øksendal. *Stochastic Differential Equations.* Springer-Verlag, Berlin, fifth edition, 2003.

22. R. G. Pinsky. *Positive Harmonic Functions and Diffusion.* Cambridge University Press, Cambridge, 1995.

23. P. E. Protter. *Stochastic Integration and Differential Equations.* Springer-Verlag, Heidelberg, second edition, 2005.

24. S. M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, second edition, 1996.

25. S. M. Ross. *A First Course in Probability.* Prentice Hall, Englewood Cliffs, NJ, seventh edition, 2005.

26. W. Rudin. *Real and Complex Analysis.* McGraw-Hill, New York, third edition, 1987.

27. I. G. Shevtsova. Sharpening of the upper bound of the absolute constant in the Berry–Esséen inequality. *Theory of Probability and Its Applications*, 51(3):549–553, 2007.

28. D. Williams. *Probability with Martingales.* Cambridge University Press, Cambridge, 1991.