1
Stochastic Processes

James R. Cruise, Ostap O. Hryniv, and Andrew R. Wade

1.1
Introduction

Basic probability theory deals, among other things, with *random variables* and their properties. A random variable is the mathematical abstraction of the following concept: we make a measurement on some physical system, subject to randomness or uncertainty, and observe the value. We can, for example, construct a mathematical model for the system and try to predict the behavior of our random observable, perhaps through its distribution, or at least its average value (mean). Even in the simplest applications, however, we are confronted by systems that change over time. Now we do not have a single random variable, but a family of random variables. The nature of the physical system that we are modeling determines the structure of dependencies of the variables.

A *stochastic* (or random) *process* is the mathematical abstraction of these systems that change randomly over time. Formally, a stochastic process is a family of random variables $(X_t)_{t\in T}$, where T is some index set representing time. The two main examples are $T = \{0, 1, 2, \dots\}$ (*discrete* time) and $T = [0, \infty)$ (*continuous* time); different applications will favor one or other of these. Interesting classes of processes are obtained by imposing additional structure on the family X_t, as we shall see.

The aim of this chapter is to give a tour of some of the highlights of stochastic process theory and its applications in the physical sciences. In line with the intentions of this volume, our emphasis is on *tools*. However, the combination of a powerful tool and an unsteady grip is a hazardous one, so we have attempted to maintain mathematical accuracy. For reasons of space, the presentation is necessarily concise. While we cover several important topics, we omit many more. We include references for further reading on the topics that we do cover throughout the text and in Section 1.8. The tools that we exhibit include *generating functions* and other transforms, and *renewal structure*, including the Markov property, which can be viewed loosely as a notion of statistical self-similarity.

In the next section, we discuss some of the tools that we will use, with some examples. The basic notions of probability

Mathematical Tools for Physicists, Second Edition. Edited by Michael Grinfeld.

theory that we use are summarized in Section 1.A.

1.2 Generating Functions and Integral Transforms

1.2.1 Generating Functions

Given a sequence $(a_k)_{k\geq 0}$ of real numbers, the function

$$G(s) = G_a(s) = \sum_{k\geq 0} a_k s^k \qquad (1.1)$$

is called the *generating function* of $(a_k)_{k\geq 0}$. When $G_a(s)$ is finite for some $s \neq 0$, the series (1.1) converges in the disc $\{z \in \mathbb{C} : |z| < |s|\}$ and thus its coefficients a_k can be recovered via differentiation

$$a_k = \frac{1}{k!}\Big(\frac{\mathrm{d}}{\mathrm{d}s}\Big)^k G_a(s)\,\Big|_{s=0}, \qquad (1.2)$$

or using the Cauchy integral formula

$$a_k = \frac{1}{2\pi i}\oint_{|z|=r} \frac{G_a(z)}{z^{k+1}}\mathrm{d}z, \quad (i^2 = -1), \quad (1.3)$$

with properly chosen $r > 0$. This observation is often referred to as the *uniqueness property*: if two generating functions, say $G_a(s)$ and $G_b(s)$, are finite and coincide in some open neighborhood of the origin, then $a_k = b_k$ for all $k \geq 0$. In particular, one can identify the sequence from its generating function.

The generating function is one of many transforms very useful in applications. We will discuss some further examples in Section 1.2.3.

Example 1.1 Imagine one needs to pay the sum of n pence using only one pence and two pence coins. In how many ways can this be done?[1)]

(a) If the order matters, that is, when $1+2$ and $2+1$ are two distinct ways of paying 3 pence, the question is about counting the number a_n of monomer/dimer configurations on the interval of length n. One easily sees that $a_0 = a_1 = 1$, $a_2 = 2$, and in general $a_n = a_{n-1} + a_{n-2}$, $n \geq 2$, because the leftmost monomer can be extended to a full configuration in a_{n-1} ways, whereas the leftmost dimer is compatible with exactly a_{n-2} configurations on the remainder of the interval. A straightforward computation using the recurrence relation above now gives

$$\begin{aligned} G_a(s) &= 1 + s + \sum_{n\geq 2}\big(a_{n-1} + a_{n-2}\big)s^n \\ &= 1 + sG_a(s) + s^2 G_a(s), \end{aligned}$$

so that $G_a(s) = (1 - s - s^2)^{-1}$. The coefficient a_n can now be recovered using (1.2), (1.3), or partial fractions and then power series expansion.

(b) If the order does not matter, that is, when $1+2$ and $2+1$ are regarded as identical ways of paying 3 pence, the question above boils down to calculating the number b_n of nonnegative integer solutions to the equation $n_1 + 2n_2 = n$. One easily sees that $b_0 = b_1 = 1$, $b_2 = b_3 = 2$, and a straightforward induction shows that b_n is the coefficient of s^k in the product

1) Early use of generating functions was often in this enumerative vein, going back to de Moivre and exemplified by Laplace in his *Théorie analytique des probabilités* of 1812. The full power of generating functions in the theory of stochastic processes emerged later with work of Pólya, Feller, and others.

$$\left(1+s+s^2+s^3+s^4+\cdots\right) \times \left(1+s^2+s^4+s^6+\cdots\right).$$

In other words, $G_b(s)=\sum_{n\geq 0} b_n s^n = (1-s)^{-1}(1-s^2)^{-1}$.

In this chapter, we will make use of generating functions for various sequences with probabilistic meaning. In particular, given a $\mathbb{Z}_+$-valued[2] random variable X, we can consider the corresponding *probability generating function*, which is the generating function of the sequence (p_k), where $p_k=\mathbb{P}[X=k]$, $k\in\mathbb{Z}_+$, describes the probability mass function of X. Thus, the probability generating function of X is given by (writing $\mathbb{E}$ for expectation: see Section 1.A)

$$G(s)=G_X(s)=\sum_{k\geq 0} s^k \mathbb{P}[X=k]=\mathbb{E}[s^X]. \tag{1.4}$$

Example 1.2

(a) If $Y\sim \mathrm{Be}(p)$ and $X\sim \mathrm{Bin}(n,p)$ (see Example 1.22), then $G_Y(s)=(1-p)+ps$ and, by the binomial theorem, $G_X(s)=\sum_{k\geq 0}\binom{n}{k}(sp)^k(1-p)^{n-k}=((1-p)+ps)^n$.

(b) If $X\sim \mathrm{Po}(\lambda)$ (see Example 1.23), then Taylor's formula implies

$$G_X(s)=\sum_{k\geq 0} s^k \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}=\mathrm{e}^{\lambda(s-1)}.$$

Notice that $G_X(1)=\mathbb{P}[X<\infty]$, and thus for $|s|\leq 1$, $|G_X(s)|\leq G_X(1)\leq 1$, implying that the power series $G_X(s)$ can be differentiated in the disk $|s|<1$ any number of times. As a result,

$$G_X^{(k)}(s)=\left(\frac{\mathrm{d}}{\mathrm{d}s}\right)^k G_X(s)=\mathbb{E}[(X)_k s^{X-k}],\quad |s|<1,$$

where $(x)_k=\frac{x!}{(x-k)!}=x(x-1)\cdots(x-k+1)$. Taking the limit $s\nearrow 1$ we obtain the kth *factorial moment* of X, $\mathbb{E}[(X)_k]=G_X^{(k)}(1_-)$, where the last expression denotes the value of the kth left derivative of $G_X(\cdot)$ at 1.

Example 1.3 If $X\sim \mathrm{Po}(\lambda)$, from Example 1.2b, we deduce $\mathbb{E}[(X)_k]=\lambda^k$.

Theorem 1.1 *If X and Y are independent $\mathbb{Z}_+$-valued random variables, then the sum $Z=X+Y$ has generating function $G_Z(s)=G_X(s)G_Y(s)$.*

Proof. As X and Y are independent, so are s^X and s^Y; consequently, the expectation factorizes, $\mathbb{E}\left[s^X s^Y\right]=\mathbb{E}\left[s^X\right]\mathbb{E}\left[s^Y\right]$, by Theorem 1.18.

Example 1.4

(a) If $X\sim \mathrm{Po}(\lambda)$ and $Y\sim \mathrm{Po}(\mu)$ are independent, then $X+Y\sim \mathrm{Po}(\lambda+\mu)$. Indeed, by Theorem 1.1, $G_{X+Y}(s)=\mathrm{e}^{(\lambda+\mu)(s-1)}$, which is the generating function of the $\mathrm{Po}(\lambda+\mu)$ distribution; the result now follows by uniqueness. This fact is known as the *additive* property of Poisson distributions.

(b) If $\mathbb{Z}_+$-valued random variables $X_1,\ldots,X_n$ are independent and identically distributed (i.i.d.), then $S_n=X_1+\cdots+X_n$ has generating function $G_{S_n}(s)=(G_X(s))^n$. If $S_n\sim \mathrm{Bin}(n,p)$, then S_n is a sum of n independent Bernoulli variables with parameter p, and so $G_{S_n}(s)=((1-p)+ps)^n$ (see Example 1.22).

The following example takes this idea further.

Lemma 1.1 *Let $X_1,X_2,\ldots$ be i.i.d. $\mathbb{Z}_+$-valued random variables, and let*

2) Throughout we use the notation $\mathbb{Z}_+=\{0,1,2,\ldots\}$ and $\mathbb{N}=\{1,2,\ldots\}$.

N be a $\mathbb{Z}_+$-valued random variable independent of the X_i. Then the random sum $S_N = X_1 + \cdots + X_N$ has generating function $G_{S_N}(s) = G_N\big(G_X(s)\big)$.

Proof. The claim follows from the partition theorem for expectations (see Section 1.A):

$$\begin{aligned}\mathbb{E}[s^{S_N}] &= \sum_{n\geq 0} \mathbb{E}[s^{S_N} \mid N = n]\mathbb{P}[N = n] \\ &= \sum_{n\geq 0} (G_X(s))^n \mathbb{P}[N = n].\end{aligned}$$

Example 1.5 If $(X_k)_{k\geq 1}$ are independent Be(p) random variables and if $N \sim \text{Po}(\lambda)$ is independent of $(X_k)_{k\geq 1}$, then

$$G_{S_N}(s) = G_N(G_X(s)) = \mathrm{e}^{\lambda(G_X(s)-1)} = \mathrm{e}^{\lambda p(s-1)},$$

that is, $S_N \sim \text{Po}(\lambda p)$. This result has the following important interpretation: if each of $N \sim \text{Po}(\lambda)$ objects is independently selected with probability p, then the sample contains $S_N \sim \text{Po}(\lambda p)$ objects. This fact is known as the *thinning* property of Poisson distributions.

A further important application of Lemma 1.1 is discussed in Section 1.2.2.

Example 1.6 [Renewals] A diligent janitor replaces a light bulb on the same day as it burns out. Suppose the first bulb is put in on day 0, and let X_i be the lifetime of the ith bulb. Suppose that the X_i are i.i.d. random variables with values in $\mathbb{N}$ and common generating function $G_f(s)$. Define $r_n = \mathbb{P}$[a bulb was replaced on day n] and $f_k = \mathbb{P}$[the first bulb was replaced on day k]. Then $r_0 = 1, f_0 = 0$, and for $n \geq 1$,

$$r_n = f_1 r_{n-1} + f_2 r_{n-2} + \cdots + f_n r_0 = \sum_{k=1}^{n} f_k r_{n-k}. \tag{1.5}$$

Proceeding as in Example 1.1, we deduce $G_r(s) = 1 + G_f(s)\,G_r(s)$ for all $|s| \leq 1$, so that

$$G_r(s) = \frac{1}{1 - G_f(s)}. \tag{1.6}$$

Remark 1.1 *The ideas behind Example 1.6 have a vast area of applicability. For example, the hitting probabilities of discrete-time Markov chains have property similar to the* Ornstein–Zernike relation (1.5)*, see, for example, Lemma 1.3.*

Notice also that by repeatedly expanding each r_i on the right-hand side of (1.5)*, one can rewrite the latter as*

$$r_n = \sum_{\ell\geq 1} \sum_{\{k_1,k_2,\ldots,k_\ell\}} \prod_{j=1}^{\ell} f_{k_j}, \tag{1.7}$$

where the middle sum runs over all decompositions of the integer n into ℓ positive integer parts $k_1, k_2, \ldots, k_\ell$ (cf. Example 1.1a). The decomposition (1.7) *is an example of the so-called* polymer representation*; it arises in many models of statistical mechanics.*

The *convolution* of sequences $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ is $(c_n)_{n\geq 0}$ given by

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}, \quad (n \geq 0); \tag{1.8}$$

we write $c = a \star b$. A key property of convolutions is as follows.

Theorem 1.2 (Convolution theorem) *If $c = a \star b$, then the associated generating functions $G_c(s)$, $G_a(s)$, and $G_b(s)$ satisfy $G_c(s) = G_a(s)G_b(s)$.*

Proof. Compare the coefficients of s^n on both sides of the equality.

The convolution appears very often in probability theory because the probability

mass function of the sum $X + Y$ of independent variables X and Y is the convolution of their respective probability mass functions. In other words, Theorem 1.1 is a particular case of Theorem 1.2.

Remark 1.2 *The sequence* $(b_n)_{n\geq 0}$ *in Example 1.1b is a convolution of the sequence* $1, 1, 1, 1, \ldots$ *and the sequence* $1, 0, 1, 0, 1, 0, \ldots$.

The uniqueness property of generating functions affords them an important role in studying convergence of probability distributions.

Theorem 1.3 *For every fixed* $n \geq 1$, *let the sequence* $a_{n,0}, a_{n,1}, \ldots$ *be a probability distribution on* $\mathbb{Z}_+$, *that is,* $a_{n,k} \geq 0$ *and* $\sum_{k\geq 0} a_{n,k} = 1$. *Moreover, let* $G_n(s) = \sum_{k\geq 0} a_{n,k} s^k$ *be the generating function of the sequence* $(a_{n,k})_{k\geq 0}$. *Then*

$$\lim_{n\to\infty} a_{n,k} = a_k, \text{ for all } k \geq 0 \iff \lim_{n\to\infty} G_n(s) = G(s), \text{ for all } s \in [0, 1),$$

where $G(s)$ *is the generating function of the limiting sequence* $(a_k)_{k\geq 0}$.

Example 1.7 [Law of rare events] Let $(X_n)_{n\geq 1}$ be random variables such that $X_n \sim \text{Bin}(n, p_n)$. If $n \cdot p_n \to \lambda$ as $n \to \infty$, then the distribution of X_n converges to that of $X \sim \text{Po}(\lambda)$. Indeed, for every fixed $s \in [0, 1)$, we have, as $n \to \infty$,

$$G_{X_n}(s) = (1 + p_n(s-1))^n \to \exp\{\lambda(s-1)\},$$

which is the generating function of a $\text{Po}(\lambda)$ random variable.

1.2.2 Example: Branching Processes

Let $(Z_{n,k})$, $n \in \mathbb{N}$, $k \in \mathbb{N}$, be a family of i.i.d. $\mathbb{Z}_+$-valued random variables with common probability mass function $(p_k)_{k\geq 0}$ and finite mean. The corresponding *branching process* $(Z_n)_{n\geq 0}$ is defined via $Z_0 = 1$, and, for $n \geq 1$,

$$Z_n = Z_{n,1} + Z_{n,2} + \cdots + Z_{n,Z_{n-1}}, \qquad (1.9)$$

where an empty sum is interpreted as zero.

The interpretation is that Z_n is the number of individuals in the nth generation of a population that evolves via a sequence of such generations, in which each individual produces offspring according to $(p_k)_{k\geq 0}$, independently of all other individuals in the same generation; generation $n + 1$ consists of all the offspring of individuals in generation n. The process starts (generation 0) with a single individual and the population persists as long as the generations are successful in producing offspring, or until *extinction* occurs. Branching processes appear naturally when modeling chain reactions, growth of bacteria, epidemics, and other similar phenomena[3]: a crucial characteristic of the process is the probability of extinction.

Let $\varphi_n(s) := \mathbb{E}[s^{Z_n}]$ be the generating function of Z_n; for simplicity, we write $\varphi(s)$ instead of $\varphi_1(s) = \mathbb{E}[s^{Z_1}]$. Then $\varphi_0(s) = s$, and a straightforward induction based on Lemma 1.1 implies

$$\varphi_n(s) = \varphi_{n-1}(\varphi(s)), \quad (n > 1). \qquad (1.10)$$

Equation (1.10) can be used to determine the distribution of Z_n for any $n \geq 0$. In particular, one easily deduces that $\mathbb{E}[Z_n] = m^n$, where $m = \mathbb{E}[Z_1]$ is the expected number of offspring of a single individual.

3) The simplest branching processes, as discussed here, are known as *Galton–Watson* processes after F. Galton and H. Watson's work on the propagation of human surnames; work on branching processes in the context of nuclear fission, by S. Ulam and others, emerged out of the Manhattan project.

The long-term behavior of a branching process is determined by the expected value m: the process can be *subcritical* ($m < 1$), *critical* ($m = 1$), or *supercritical* ($m > 1$).

Remark 1.3 *By Markov's inequality (see Section 1.A),* $\mathbb{P}[Z_n > 0] = \mathbb{P}[Z_n \geq 1] \leq \mathbb{E}[Z_n] = m^n$. *Hence, in the* subcritical *case,* $\mathbb{P}[Z_n > 0] \to 0$ *as* $n \to \infty$ *(i.e.,* $Z_n \to 0$ in probability*). Moreover, the average* total population *in this case is finite, because* $\mathbb{E}[\sum_{n\geq 0} Z_n] = \sum_{n\geq 0} m^n = (1-m)^{-1} < \infty$. *It follows that, with probability 1,* $\sum_{n\geq 0} Z_n < \infty$, *which entails* $Z_n \to 0$ almost surely. *This last statement can also be deduced from the fact that* $\sum_{n\geq 0} \mathbb{P}[Z_n > 0] < \infty$ *using the* Borel–Cantelli lemma *(see, e.g.,* [1, 2]*).*

Extinction is the event $\mathcal{E} = \cup_{n=1}^{\infty} \{Z_n = 0\}$. As $\{Z_n = 0\} \subseteq \{Z_{n+1} = 0\}$ for all $n \geq 0$, the *extinction probability* $\rho = \mathbb{P}[\mathcal{E}]$ is well defined (by continuity of probability measure: see Section 1.A) via $\rho = \lim_{n\to\infty} \mathbb{P}[Z_n = 0]$, where $\mathbb{P}[Z_n = 0] = \varphi_n(0)$ is the probability of extinction *before* the $(n+1)$th generation.

Theorem 1.4 *If* $0 < p_0 < 1$, *then the extinction probability* ρ *is given by the smallest positive solution to the equation*

$$s = \varphi(s). \tag{1.11}$$

In particular, if $m = \mathbb{E}[Z] \leq 1$, *then* $\rho = 1$; *otherwise, we have* $0 < \rho < 1$.

Remark 1.4 *The relation* (1.11) *has a clear probabilistic sense. Indeed, by independence,* $\mathbb{P}[\mathcal{E} \mid Z_1 = k] = \rho^k$ *(as extinction only occurs if each of the independent branching processes associated with the* k *individuals dies out). Then, by the law of total probability (see Section 1.A), we get*[4)]

$$\rho = \mathbb{P}[\mathcal{E}] = \sum_{k\geq 0} \mathbb{P}[\mathcal{E} \mid Z_1 = k]\mathbb{P}[Z_1 = k] = \sum_{k\geq 0} \rho^k \mathbb{P}[Z_1 = k] = \varphi(\rho).$$

Notice that Theorem 1.4 characterizes the extinction probability without the necessity to compute $\varphi_n(\cdot)$. The excluded values $p_0 \in \{0, 1\}$ are trivial: if $p_0 = 0$, then $Z_n \geq 1$ for all $n \geq 0$ so that $\rho = 0$; if $p_0 = 1$, then $\mathbb{P}[Z_1 = 0] = \rho = 1$.

Proof of Theorem 1.4 Denote $\rho_n = \mathbb{P}[Z_n = 0] = \varphi_n(0)$. By continuity and strict monotonicity of the generating function $\varphi(\cdot)$, we have (recall (1.10))

$$0 < \rho_1 = \varphi(0) < \rho_2 = \varphi(\rho_1) < \cdots < 1 = \varphi(1),$$

so that $\rho_n \nearrow \rho \in (0, 1]$ with $\rho = \varphi(\rho)$.

Now if $\overline{\rho}$ is another fixed point of $\varphi(\cdot)$ in $[0, 1]$, that is, $\overline{\rho} = \varphi(\overline{\rho})$, then, by induction,

$$0 < \rho_1 = \varphi(0) < \rho_2 < \cdots < \varphi(\overline{\rho}) = \overline{\rho},$$

so that $\rho = \lim_{n\to\infty} \rho_n \leq \overline{\rho}$, that is, ρ is the smallest positive solution to (1.11).

Finally, by continuity and convexity of $\varphi(\cdot)$ together with the fact $\varphi(1) = 1$, the condition $m = \varphi'(1) \leq 1$ implies $\rho = 1$ and the condition $m = \varphi'(1) > 1$ implies that ρ is the unique solution in $(0, 1)$ to the fixed point equation (1.11).

We thus see that the branching process exhibits a phase transition: in the subcritical or critical regimes ($m \leq 1$), the process dies out with probability 1, whereas in the supercritical case ($m > 1$) it survives forever with positive probability $1 - \rho$.

4) In Section 1.3, we will see this calculation as exploiting the *Markov property* of the branching process.

1.2.3 Other Transforms

1.2.3.1 Moment Generating Functions

The *moment generating function* of a real-valued random variable X is defined by $M_X(t) = \mathbb{E}[e^{tX}]$. When finite for t in some neighborhood of 0, $M_X(t)$ behaves similarly to the generating function $G_X(s)$ in that it possesses the uniqueness property (identifying the corresponding distribution), maps convolutions (i.e., distributions of sums of independent variables) into products, and can be used to establish convergence in distribution.

If X is $\mathbb{Z}_+$-valued and its generating function $G_X(s)$ is finite for some $s > 1$, then $M_X(t)$ is also finite for some $t \neq 0$, and the two are related by $M_X(t) = G_X(e^t)$. For example, if $X \sim \mathrm{Bin}(n,p)$, then $M_X(t) = (1 + p(e^t - 1))^n$ (see Example 1.2a). The terminology arises from the fact that if $M_X(t)$ is differentiable at 0, then the kth derivative of $M_X(t)$ evaluated at 0 gives $\mathbb{E}[X^k]$, the kth *moment* of X.

Example 1.8 In the case of a continuous distribution X with probability density f, the moment generating function $M_X(t)$ becomes the integral transform $\int e^{tx} f(x)\mathrm{d}x$.

(a) Let $\lambda > 0$. If X has density $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, and 0 elsewhere, then X has the *exponential distribution* with parameter λ and

$$M_X(t) = \int_0^\infty \lambda e^{-(\lambda - t)x}\mathrm{d}x = \frac{\lambda}{\lambda - t}, \quad (t < \lambda).$$

(b) If X has density $f(x) = e^{-x^2/2}/\sqrt{2\pi}$, $x \in \mathbb{R}$, then X has the *standard normal* $\mathcal{N}(0,1)$ or *Gaussian* distribution and

$$M_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t^2/2} e^{-(x-t)^2/2}\mathrm{d}x = e^{t^2/2}, \quad (t \in \mathbb{R}).$$

The *normal distribution* was in part so named[5] for its ubiquity in real data. It is also very common in probability and mathematical statistics, owing to a large extent to results of the following kind.

Theorem 1.5 (de Moivre – Laplace central limit theorem) *Let $X_n \sim \mathrm{Bin}(n,p)$ with fixed $p \in (0,1)$. Denote $X_n^* = (X_n - \mathbb{E}[X_n])/\sqrt{\mathbb{V}\mathrm{ar}[X_n]}$. Then for any $t \in \mathbb{R}$,*

$$M_{X_n^*}(t) \to e^{t^2/2}, \text{ as } n \to \infty;$$

in other words, the distribution of X_n^* converges to that of $\mathcal{N}(0,1)$.

Proof. Recall that X_n can be written as a sum $Y_1 + Y_2 + \cdots + Y_n$ of independent $\mathrm{Be}(p)$ random variables. In particular, $\mathbb{E}[X_n] = np$ and $\mathbb{V}\mathrm{ar}[X_n] = np(1-p)$ (see Example 1.25). Using the simple relation $M_{aZ+b}(t) = e^{bt} M_Z(at)$, we deduce that the random variable $\widehat{Y} = (Y - p)/\sqrt{n}$ has the moment generating function (with fixed t)

$$e^{-tp/\sqrt{n}}\left(1 + p\left(e^{t/\sqrt{n}} - 1\right)\right) = 1 + \frac{t^2}{2n}p(1-p) + O\left(\frac{t^3}{\sqrt{n}}\right).$$

5) The term *normal* (in the sense of "usual") was apparently attached to the distribution by Francis Galton and others and popularized by Karl Pearson. The distribution arose in the work of Gauss and Laplace on least squares and errors of measurement, and also in Maxwell's work on statistical physics. Perhaps its first tentative appearance, however, is in the work of Abraham de Moivre for whom Theorem 1.5 is named. See [3] for a discussion.

Noticing that $M_{X_n^*}(t) = (M_{\hat{Y}}(t/\sqrt{p(1-p)}))^n \to \mathrm{e}^{t^2/2}$ as $n \to \infty$, we deduce the result.

One other application of moment generating functions that we will see is to *large deviations*: see Section 1.4.4.

1.2.3.2 Laplace Transforms

In general, $M_X(t)$ might be infinite for all $t \neq 0$. However, for nonnegative variables $X \geq 0$, we have $M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] \leq 1$, for all $t \leq 0$; in particular, $M_X(t)$ is an analytic function at every point of the complex plane with negative real part. In this case, $M_X(t)$ behaves very similarly to generating functions and inherits the main properties described above. In such a situation, the function $M_X(t)$ (or, sometimes $M_X(-t)$) is called the *Laplace transform* of the variable X. See Chapter 15 in the present volume for background on Laplace transforms.

1.2.3.3 Characteristic Functions

Unlike the moment generating function $M_X(t)$, which might be infinite for real $t \neq 0$, the *characteristic function* $\psi_X(t) = \mathbb{E}[\mathrm{e}^{\mathrm{i}tX}]$ (where $\mathrm{i}^2 = -1$) always exists and uniquely identifies the distribution, hence the name.[6] The characteristic functions inherit all nice properties of (moment) generating functions, though inverting them is not always straightforward.

Characteristic functions are the standard tool of choice for proving results such as the following generalization of Theorem 1.5. The proof is similar to that of the previous theorem, based on a Taylor-type formula: if $\mathbb{E}[X^{2n}] < \infty$, then

$$\psi_X(t) = \sum_{\ell=0}^{2n} \frac{(\mathrm{i}t)^\ell}{\ell!} \mathbb{E}[X^\ell] + o(t^{2n}).$$

Theorem 1.6 (Central limit theorem) *Let $X_n = Y_1 + Y_2 + \cdots + Y_n$, where Y_i are i.i.d. random variables with $\mathbb{E}[Y_i^2] < \infty$. Then, as $n \to \infty$, the distribution of $X_n^* = (X_n - \mathbb{E}[X_n])/\sqrt{\mathbb{V}\mathrm{ar}[X_n]}$ converges to the standard normal, $\mathcal{N}(0,1)$.*

1.3 Markov Chains in Discrete Time

1.3.1 What is a Markov Chain?

Our tour begins with stochastic processes in *discrete time*. Here, we will write our process as $X_0, X_1, X_2, \ldots$. A fundamental class is constituted by the *Markov processes* in which, roughly speaking, *given the present, the future is independent of the past*. In this section, we treat the case where the X_n take values in a *discrete* (i.e., finite or countably infinite) state space S. In this case, the general term Markov process is often specialized to a *Markov chain*, although the usage is not universally consistent.

The process $X = (X_n)$ taking values in the discrete set S satisfies the *Markov property* if, for any n and any $i, j, \in S$,[7]

$$\mathbb{P}[X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0] = p_{ij}, \tag{1.12}$$

for all previous histories $i_0, \ldots, i_{n-1} \in S$.

The $p_{ij} = \mathbb{P}[X_{n+1} = j \mid X_n = i]$ are the *one-step transition probabilities* for X, and they satisfy the obvious conditions $p_{ij} \geq 0$ for all i, j, and $\sum_{j \in S} p_{ij} = 1$ for all i. It is convenient to arrange the p_{ij} as a matrix $P = (p_{ij})_{i,j \in S}$ with nonnegative entries and

6) The term *characteristic function* is traditional in the probabilistic context for what elsewhere might be called the *Fourier transform*: see Chapter 15 of the present volume.

7) In (1.12) and elsewhere, we indicate intersections of events by commas for readability.

whose rows all sum to 1: these properties define a *stochastic matrix*. The Markov property specifies the step-by-step evolution of the Markov chain. One can imagine a particle moving at random on S, from state i selecting its next location according to distribution $(p_{ij})_{j\in S}$. It is an exercise in conditional probability to deduce from (1.12) that

$$\mathbb{P}[X_1 = i_1, \dots, X_n = i_n \mid X_0 = i_0] = p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.$$

It may be that X_0 is itself random, in which case, in order to assign a probability to any particular (finite) sequence of moves for the particle, in addition to the Markov property, we also need to know where the particle starts: this is specified by the *initial distribution* $\mathbb{P}[X_0 = i] = w_i$, $i \in S$.

Now, to compute the probability of getting from i to k in *two* steps, we sum over the j-partition to get

$$\mathbb{P}[X_2 = k \mid X_0 = i] = \sum_{j\in S} \mathbb{P}[X_1 = j, X_2 = k \mid X_0 = i] = \sum_{j\in S} p_{ij} p_{jk} = (P^2)_{ik},$$

the (i,k) entry in $P^2 = P \cdot P$; this *matrix multiplication* yields the *two-step transition probability*. More generally, the n-step transition probabilities are

$$p_{ij}^{(n)} := \mathbb{P}[X_n = j \mid X_0 = i] = (P^n)_{ij}.$$

A similar argument shows the following:

$$p_{ij}^{(n+m)} = \sum_{k\in S} p_{ik}^{(n)} p_{kj}^{(m)}. \tag{1.13}$$

Remark 1.5 *We restrict ourselves to the case of* time-homogeneous *Markov chains, by stipulating that (1.12) should hold simultaneously for all n. One may relax this, and allow p_{ij} to depend also on n.*

Remark 1.6 *An equivalent definition of a Markov chain is as a* randomized dynamical system: $X_{n+1} = f(X_n; U_{n+1})$ *where* $f : S \times [0,1] \to S$ *is a fixed update function, and* $U_1, U_2, \dots$ *are independent* $U[0,1]$ *random variables.*[8] Monte Carlo simulation *of a Markov chain uses such a scheme.*

We make a final remark on notation: for p_{ij}, and similar expressions, we sometimes write $p_{i,j}$ if otherwise the subscripts may be ambiguous.

1.3.2 Some Examples

Example 1.9 [The Ehrenfest model] In 1907, Ehrenfest and Ehrenfest [4] introduced this simple model of diffusion. There are N particles in a container that has two chambers separated by a permeable partition. At each step, a particle is chosen uniformly at random and moved across the partition. The state of the Markov chain at each time will be the number of particles in the first chamber, say, so $S = \{0, \dots, N\}$.

The one-step transition probabilities are, for $i \in \{1, \dots, N-1\}$,

$$p_{i,i+1} = \frac{N-i}{N}, \quad p_{i,i-1} = \frac{i}{N},$$

and $p_{0,1} = p_{N,N-1} = 1$.

After a long time, what is the distribution of the particles? See Section 1.3.6.

Example 1.10 [One-dimensional simple random walk] A particle moves at random on the state space $S = \mathbb{Z}_+$. From position $i \neq 0$, the particle jumps one step to the left with probability p_i and one step to the right with probability $1 - p_i$. With partial reflection at 0, we can describe this *random*

8) That is, uniform on $[0,1]$, having density $f(x) = 1$ for $x \in [0,1]$ and $f(x) = 0$ elsewhere.

walk by a Markov chain with one-step transition probabilities $p_{0,0} = p_0 \in (0,1)$, $p_{0,1} = q_0 := 1 - p_0$, and for $i \geq 1$,

$$p_{i,i-1} = p_i \in (0,1), \quad p_{i,i+1} = q_i := 1 - p_i.$$

1.3.3 Stationary Distribution

We use the compact notation $\mathbb{P}_i$ for the (conditional) probability associated with the Markov chain started from state $i \in S$, that is, $\mathbb{P}_i[\,\cdot\,] = \mathbb{P}[\,\cdot \mid X_0 = i]$. More generally, if $w = (w_i)_{i\in S}$ is a distribution on S (i.e., $w_i \geq 0$, $\sum_i w_i = 1$), then we write $\mathbb{P}_w$ for the Markov chain started from the initial distribution w, that is, $\mathbb{P}_w[\,\cdot\,] = \sum_i w_i \mathbb{P}_i[\,\cdot\,]$.

A distribution $\pi = (\pi_i)_{i\in S}$ is a *stationary distribution* for a Markov chain X if

$$\mathbb{P}_\pi[X_1 = i] = \pi_i, \text{ for all } i \in S. \tag{1.14}$$

Viewing a stationary distribution π as a row vector, (1.14) is equivalent to the matrix-vector equation $\pi P = \pi$, that is, π is a left eigenvector of P corresponding to the eigenvalue 1. The nomenclature arises from the fact that (1.14) implies that $\mathbb{P}_\pi[X_n = i] = \pi_i$ for *all* times n, so the distribution of the Markov chain started according to π is stationary in time.

Example 1.11 [A three-state chain] Consider a Markov chain (X_n) with the state space $\{1,2,3\}$ and transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{1.15}$$

We look for a stationary distribution $\pi = (\pi_1, \pi_2, \pi_3)$. Now $\pi P = \pi$ with the fact that $\pi_1 + \pi_2 + \pi_3 = 1$ gives a system of equations with unique solution $\pi = (\frac{2}{7}, \frac{3}{7}, \frac{2}{7})$.

A Markov chain with transition probabilities p_{ij} is *reversible* with respect to a distribution $\pi = (\pi_i)_{i\in S}$ if the *detailed balance equations* hold:

$$\pi_i p_{ij} = \pi_j p_{ji}, \text{ for all } i, j \in S. \tag{1.16}$$

Not every Markov chain is reversible. Any distribution π satisfying (1.16) is necessarily a stationary distribution, because then, for all j, $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j$. If the chain *is* reversible, then the system of equations (1.16) is often simpler to solve than the equations $\pi P = \pi$: see Example 1.12. The "physical" interpretation of reversibility is that, in equilibrium, the Markov chain is statistically indistinguishable from a copy of the chain running backward in time.

Example 1.12 [Random walk] Consider Example 1.10. We seek a solution $\pi = (\pi_i)$ to (1.16), which now reads $\pi_i q_i = \pi_{i+1} p_{i+1}$ for all $i \geq 0$. The solution is $\pi_i = \pi_0 \prod_{j=0}^{i-1} (q_j / p_{j+1})$. This describes a proper distribution if $\sum_i \pi_i = 1$, that is, if

$$\sum_{i=0}^{\infty} \prod_{j=0}^{i-1} \frac{q_j}{p_{j+1}} < \infty. \tag{1.17}$$

If (1.17) holds, then

$$\pi_i = \frac{\prod_{j=0}^{i-1} \frac{q_j}{p_{j+1}}}{\sum_{i=0}^{\infty} \prod_{j=0}^{i-1} \frac{q_j}{p_{j+1}}}.$$

For example, if $p_i = p \in (0,1)$ for all i, then (1.17) holds if and only if $p > 1/2$, in which case $\pi_i = (q/(1-2p))(q/p)^i$, where $q = 1 - p$, an exponentially decaying stationary distribution.

1.3.4 The Strong Markov Property

One way of stating the Markov property is to say that $(X_n)_{n\geq m}$, conditional on

$\{X_m = i\}$, is distributed as the Markov chain $(X_n)_{n\geq 0}$ with initial state $X_0 = i$. It is often desirable to extend such a statement from deterministic times m to *random* times T. An important class of random times are the *first passage times*,[9]

$$T_i := \min\{n \geq 1 : X_n = i\},\ i \in S. \quad (1.18)$$

The Markov property cannot hold at every random time. For instance, if $T' = T_i - 1$, then the first transition of the process $X_{T'}, X_{T'+1}, \ldots$ is always from $X_{T'}$ to $X_{T'+1} = i$, regardless of the original transition matrix.

The following *strong Markov property* clarifies these issues. A random time $T \in \mathbb{Z}_+ \cup \{\infty\}$ is a *stopping time* with respect to (X_n) if, for any n, the event $\{T \leq n\}$ depends only on $X_0, \ldots, X_n$ (and not on the future evolution of the chain). The passage times T_i are stopping times, but T' described above is not a stopping time.

Lemma 1.2 *Suppose that T is a stopping time for (X_n). Then, given $T < \infty$ and $X_T = i$, $(X_{T+n})_{n\geq 0}$ has the same distribution as $(X_n)_{n\geq 0}$ started from $X_0 = i$.*

Sketch of proof Partition over the possible values of T. Suppose that $T = m$ and $X_T = X_m = i$; this is a condition only on $X_0, \ldots, X_m$, because T is a stopping time. Now apply the usual Markov property at the deterministic time m.

1.3.5 The One-Step Method

In problems involving Markov chains, often quantities of interest are *hitting probabilities* and *expected hitting times*. One approach to computing these is via the powerful *one-step method*, which makes essential use of the Markov property.

Recall the definition of the passage times T_i from (1.18). The *expected hitting time* of state j starting from state i is $\mathbb{E}_i[T_j]$ for $i \neq j$; if $i = j$ this is the *expected return time* to i. Also of interest is $\mathbb{P}_i[T_j < T_k]$, the probability of reaching state j before state k, starting from i. We illustrate the one-step method by some examples.

Example 1.13 [Three-state chain] We return to Example 1.11. We partition over the *first step* of the process to obtain, via the law of total probability (see Section 1.A),

$$\begin{aligned}\mathbb{P}_2[T_1 < T_3] &= \sum_{k=1}^{3} \mathbb{P}_2[\{T_1 < T_3\} \cap \{X_1 = k\}] \\ &= p_{2,1} \cdot 1 + p_{2,2} \cdot \mathbb{P}_2[T_1 < T_3] \\ &\quad + p_{2,3} \cdot 0,\end{aligned}$$

by the Markov property. This gives $\mathbb{P}_2[T_1 < T_3] = 1/2$.

What about $\mathbb{E}_2[T_1]$? Set $z_i = \mathbb{E}_i[T_1]$. Again we condition on the first step, and now use the partition theorem for expectations (see Section 1.A):

$$\begin{aligned}\mathbb{E}_2[T_1] &= 1 + \mathbb{E}_2[T_1 - 1] \\ &= 1 + \sum_{k=1}^{3} p_{2,k} \mathbb{E}_2[T_1 - 1 \mid X_1 = k].\end{aligned}$$

Now applying the Markov property at time 1, we see that $T_1 - 1$, given $X_0 \neq 1$ and $X_1 = k \neq 1$, has the same distribution as T_1 given $X_0 = k \neq 1$ in the original chain, and, in particular, has expected value z_k. On the other hand, if $X_1 = k = 1$, then $T_1 - 1 = 0$. So we get $z_2 = 1 + \frac{1}{3}z_2 + \frac{1}{3}z_3$. A similar argument starting from state 3 gives $z_3 = 1 + \frac{1}{2}z_2 + \frac{1}{2}z_3$. This system of linear equations is easily solved to give $z_2 = 5$ and $z_3 = 7$.

9) Here and elsewhere, the convention $\min \emptyset = \infty$ is in force.

Example 1.14 [Random walk; gambler's ruin] Recall Example 1.10. Fix $n \in \mathbb{N}$ and for $i \in \{1, \dots, n-1\}$, let $u_i = \mathbb{P}_i[T_n < T_0]$. The one-step method gives

$$u_i = p_i u_{i-1} + q_i u_{i+1}, \quad (1 \le i \le n-1),$$

with boundary conditions $u_0 = 0$, $u_n = 1$. The standard method to solve this system of equations is to rewrite it in terms of the differences $\Delta_i = u_{i+1} - u_i$ to get $\Delta_i = \Delta_{i-1}(p_i/q_i)$ for $1 \le i \le n-1$, which yields $\Delta_j = \Delta_0 \prod_{k=1}^{j}(p_k/q_k)$. Then $u_i = \sum_{j=0}^{i-1} \Delta_j$, using the boundary condition at 0. Using the boundary condition at n to fix Δ_0, the solution obtained is

$$u_i = \frac{\sum_{j=0}^{i-1} \prod_{k=1}^{j} \frac{p_k}{q_k}}{\sum_{j=0}^{n-1} \prod_{k=1}^{j} \frac{p_k}{q_k}}. \tag{1.19}$$

In the special case where $p_i = q_i = 1/2$ for all i, we have the elegant formula $u_i = i/n$. If we imagine that the state of the Markov chain is the wealth of a gambler with initial wealth i who plays a sequence of fair games, each time either gaining or losing a unit of wealth, $1 - u_i$ is the *ruin probability* (and u_i is the probability that the gambler makes his fortune).

1.3.6 Further Computational Methods

We present by example some additional techniques.

Example 1.15 [Matrix diagonalization] In many situations, we want to compute the n-step transition probability $p_{ij}^{(n)}$, that is, an entry in the matrix power P^n. To calculate P^n, we try to *diagonalize* P to obtain $P = T\Lambda T^{-1}$ for an invertible matrix T and a diagonal matrix Λ. The usefulness of this representation is that $P^n = T\Lambda^n T^{-1}$ and Λ^n is easy to write down, because Λ is diagonal. A sufficient condition for P to be diagonalizable is that all its eigenvalues be distinct.

Consider again the three-state chain with transition matrix given by (1.15); we have three eigenvalues, $\lambda_1, \lambda_2, \lambda_3$, say. As P is a stochastic matrix, 1 is always an eigenvalue: $\lambda_1 = 1$, say. Then because $\operatorname{tr} P = \lambda_1 + \lambda_2 + \lambda_3$ and $\det P = \lambda_1\lambda_2\lambda_3$, we find $\lambda_2 = 1/2$ and $\lambda_3 = -1/6$, say.

It follows from the diagonalized representation that

$$\begin{aligned} P^n &= \lambda_1^n U_1 + \lambda_2^n U_2 + \lambda_3^n U_3 \\ &= U_1 + \left(\frac{1}{2}\right)^n U_2 + \left(-\frac{1}{6}\right)^n U_3, \end{aligned} \tag{1.20}$$

where U_1, U_2, U_3 are 3×3 matrices to be determined. One can solve the simultaneous matrix equations arising from the cases $n \in \{0, 1, 2\}$ of (1.20) to obtain U_1, U_2, and U_3, and hence,

$$\begin{aligned} P^n = &\begin{pmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \end{pmatrix} + \left(\frac{1}{2}\right)^n \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \\ &+ \left(-\frac{1}{6}\right)^n \begin{pmatrix} \frac{3}{14} & -\frac{3}{7} & \frac{3}{14} \\ -\frac{2}{7} & \frac{4}{7} & -\frac{2}{7} \\ \frac{3}{14} & -\frac{3}{7} & \frac{3}{14} \end{pmatrix}. \end{aligned}$$

It follows that $\lim_{n\to\infty} p_{ij}^{(n)}$ exists, does not depend on i, and is equal to π_j, the component of the stationary distribution that we calculated in Example 1.11. After a long time, the chain "forgets" its starting state and approaches a stochastic equilibrium described by the stationary distribution. This is an example of a general phenomenon to which we return in Section 1.3.7.

Example 1.16 [Generating functions] We sketch the use of generating functions to

evaluate stationary distributions. Consider the Ehrenfest model of Example 1.9. Suppose that $\pi = (\pi_0, \dots, \pi_N)$ is a stationary distribution for the Markov chain, with generating function $\hat{\pi}(s) = \sum_{i=0}^{N} \pi_i s^i$. In this case, the equation $\pi P = \pi$ reads

$$\pi_{j-1} \frac{N-(j-1)}{N} + \pi_{j+1} \frac{j+1}{N} = \pi_j,$$

which is valid for all $j \in \{0, \dots, N\}$, provided we set $\pi_{-1} = \pi_{N+1} = 0$. Now multiply through by s^j and sum from $j = 0$ to N. After some algebra, we obtain

$$\hat{\pi}(s) = \frac{1-s^2}{N} \hat{\pi}'(s) + s\hat{\pi}(s),$$

so that $\mathrm{d}/\mathrm{d}s \log \hat{\pi}(s) = \hat{\pi}'(s)/\hat{\pi}(s) = N/(1+s)$. Integrating with respect to s and using the fact that $\hat{\pi}(1) = 1$, we obtain

$$\hat{\pi}(s) = \left(\frac{1+s}{2}\right)^N.$$

The binomial theorem now enables us to identify $\pi_i = 2^{-N}\binom{N}{i}$.

1.3.7 Long-term Behavior; Irreducibility; Periodicity

We saw in Example 1.15 a Markov chain for which

$$\lim_{n\to\infty} \mathbb{P}_i[X_n = j] = \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \qquad (1.21)$$

for all $i, j \in S$, where π_j is from a stationary distribution. For which Markov chains does such a result hold? There are (at least) three obstacles:

(a) There might be no solutions to $\pi P = \pi$, and hence, no right-hand side in (1.21).

(b) There might be *multiple* solutions to $\pi P = \pi$, and so no uniqueness in (1.21). For example, consider the Markov chain on the state space $\{0, 1, 2\}$ with $p_{00} = 1$, $p_{22} = 1$ (0 and 2 are *absorbing states*) and $p_{10} = p_{12} = 1/2$. Then $p_{i2}^{(n)} = i/2$ for all $n \geq 1$, that is, the limit on the left-hand side of (1.21) depends on the starting state i. Note that here $\pi = (\alpha, 0, 1-\alpha)$ is stationary for any $\alpha \in [0, 1]$.

(c) In the Ehrenfest model of Example 1.9, there is a parity effect, because $p_{00}^{(n)} = 0$ for *odd* n, for instance. This phenomenon is an example of *periodicity*, which is another obstacle to (1.21).

Cases (b) and (c) here can be dealt with after some additional concepts are introduced. A state $i \in S$ has *period* d if d is the greatest common divisor of $\{n \geq 1 : p_{ii}^{(n)} > 0\}$. For example, all states in the Ehrenfest model have period 2.

A Markov chain is *irreducible* if, for all $i, j \in S$, there exist finite m and n for which $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$, that is, it is possible to get between any two states in a finite number of steps. For the rest of this section, we will assume that we have an irreducible Markov chain. We do not discuss the case of nonirreducible (*reducible*) chains in a systematic way, but Section 1.3.10 provides an illustrative example.

For an irreducible chain, it can be shown that all states have the same period, in which case one can speak about the period of the chain itself. If all states have period 1, the chain is called *aperiodic*.

Recall the definition of T_i from (1.18): $\mathbb{E}_i[T_i]$ is the *expected return time* to i. The following result answers our question on the limiting behavior of $p_{ij}^{(n)}$.

Theorem 1.7 *For an irreducible Markov chain, the following are equivalent.*

- *There exists a unique stationary distribution* π.
- *For some* $i \in S$, $\mathbb{E}_i[T_i] < \infty$.
- *For all* $i \in S$, $\mathbb{E}_i[T_i] < \infty$.

If these conditions hold, the Markov chain is called positive recurrent. *For a positive-recurrent chain, the following hold.*

- *For all* $i \in S$, $\pi_i = 1/\mathbb{E}_i[T_i]$.
- *If the chain is aperiodic, then* $\mathbb{P}_i[X_n = j] \to \pi_j$ *for all* $i, j \in S$.

In particular, we have the following result.

Theorem 1.8 *An irreducible Markov chain on a finite state space is positive recurrent.*

Proofs of these results can be found in [5, 6], for instance.

1.3.8 Recurrence and Transience

Recall the definition of T_i from (1.18). A state $i \in S$ is called *recurrent* if $\mathbb{P}_i[T_i < \infty] = 1$ or *transient* if $\mathbb{P}_i[T_i = \infty] > 0$. A Markov chain will return infinitely often to a recurrent state, but will visit a transient state only finitely often. If a Markov chain is irreducible (see Section 1.3.7), then either all states are recurrent, or none is, and so we can speak of recurrence or transience of the chain itself.

If an irreducible chain is positive recurrent (see Theorem 1.7), then it is necessarily recurrent. A chain that is recurrent but not positive recurrent is *null recurrent*, in which case, for all i, $\mathbb{P}_i[T_i < \infty] = 1$ but $\mathbb{E}_i[T_i] = \infty$ (equivalently, it is recurrent but no stationary distribution exists). Because of Theorem 1.8, we know that to observe null recurrence or transience we must look at infinite state spaces.

Example 1.17 [One-dimensional random walk] We return to Example 1.10. Consider $\mathbb{P}_0[T_0 = \infty]$. In order for the walk to never return to 0, the first step must be to 1, and then, starting from 1, the walk must reach n before 0 for *every* $n \geq 2$. Thus

$$\begin{aligned}\mathbb{P}_0[T_0 = \infty] &= q_0 \mathbb{P}_1[T_0 = \infty] \\ &= q_0 \mathbb{P}_1[\cap_{n\geq 2}\{T_n < T_0\}].\end{aligned}$$

Note $\{T_{n+1} < T_0\} \subseteq \{T_n < T_0\}$, so the intersection here is over a decreasing sequence of events. Thus by continuity of probability measures (see Section 1.A), $\mathbb{P}_0[T_0 = \infty] = q_0 \lim_{n\to\infty} \mathbb{P}_1[T_n < T_0]$. Here $\mathbb{P}_1[T_n < T_0] = 1 - u_1$ where u_1 is given by (1.19). So we obtain

$$\mathbb{P}_0[T_0 = \infty] > 0 \text{ if and only if } \sum_{j=0}^{\infty} \prod_{k=1}^{j} \left(\frac{p_k}{q_k}\right) < \infty.$$

(For further discussion, see [7, pp. 65–71]) In particular, if $p_k = p \in (0,1)$ for all k, the walk is transient if $p < 1/2$ and recurrent if $p \geq 1/2$. The phase transition can be probed more precisely by taking $p_k = 1/2 + c/k$; in this case, the walk is transient if and only if $c < -1/4$, a result due to Harris and greatly generalized by Lamperti [8].

We give one criterion for recurrence that we will use in Section 1.4.

Lemma 1.3 *For any* $i \in S$, $\mathbb{P}_i[T_i < \infty] = 1$ *if and only if* $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$.

We give a proof via generating functions. Write $f_i^{(n)} = \mathbb{P}_i[T_i = n]$, the probability that the first return to i occurs at time n; here $f_i^{(0)} = 0$; note that $\sum_n f_i^{(n)}$ may be less than 1. Denote the corresponding generating function by $\phi_i(s) = \sum_{n=0}^{\infty} f_i^{(n)} s^n$. Also define $\psi_i(s) = \sum_{n=0}^{\infty} p_{ii}^{(n)} s^n$, where $p_{ii}^{(n)} = \mathbb{P}_i[X_n = i]$ (so $p_{ii}^{(0)} = 1$). By conditioning on the

value of T_i, the strong Markov property gives

$$p_{ii}^{(n)} = \sum_{m=0}^{n} f_i^{(m)} p_{ii}^{(n-m)}, \ (n \geq 1).$$

Treating the case $n = 0$ carefully, it follows that

$$\psi_i(s) = 1 + \sum_{n=0}^{\infty} \sum_{m=0}^{n} f_i^{(m)} p_{ii}^{(n-m)} s^n.$$

The final term here is a discrete convolution of the generating function (cf. Theorem 1.2), so we deduce the important *renewal relation*

$$\psi_i(s) = 1 + \phi_i(s)\psi_i(s). \tag{1.22}$$

Sketch of proof of Lemma 1.3 We have $\mathbb{P}_i[T_i < \infty] = \lim_{s\uparrow 1} \phi_i(s)$, and (1.22) implies that the latter limit is 1 if and only if $\lim_{s\uparrow 1} \psi_i(s) = \infty$.

1.3.9 Remarks on General State Spaces

In the case of discrete state spaces, (1.13) corresponds to the trivial matrix equation $P^{n+m} = P^n \cdot P^m$, which one could describe, rather grandly, as the *semigroup property* of matrix multiplication. More generally, (1.13) is an instance of the fundamental Chapman–Kolmogorov relation, and the connection to semigroup theory runs deep.

In a general state space, the analogue of the transition probability p_{ij} is a transition kernel $p(x; A)$ given by $p(x; A) = \mathbb{P}[X_{n+1} \in A \mid X_n = x]$. This immediately introduces technical issues that can only be addressed in the context of measure theory. We refer to [2, 9], for example.

1.3.10 Example: Bak–Sneppen and Related Models

Bak and Sneppen [10] introduced a simple stochastic model of evolution that initiated a considerable body of research by physicists and mathematicians. In the original model, N sites are arranged in a ring. Each site, corresponding to a species in the evolution model, is initially assigned an independent $U[0, 1]$ random variable representing a "fitness" value for the species. The Bak–Sneppen model is a discrete-time Markov process, where at each step the minimal fitness value and the values at the two neighboring sites are replaced by three independent $U[0, 1]$ random variables.

This process is a Markov process on the continuous state space $[0, 1]^N$, and its behavior is still not fully understood, despite a large physics literature devoted to these models: see the thesis [11] for an overview of the mathematical results.

Here we treat a much simpler model, following [12]. The state space of our process (X_n) will be the "simplex" of ranked sequence of N fitness values

$$\Delta_N := \{(x^{(1)}, \dots, x^{(N)}) \in [0, 1]^N : x^{(1)} \leq \cdots \leq x^{(n)}\}.$$

Fix a parameter $k \in \{1, \dots, N\}$. We start with N independent $U[0, 1]$ values: rank these to get X_0. Given X_n, discard the *kth-ranked* value $X_n^{(k)}$ and replace it by a new independent $U[0, 1]$ random variable; rerank to get X_{n+1}.

For example, if $k = 1$, we replace the minimal value at each step. It is natural to anticipate that X_n should approach (as $n \to \infty$) a limiting (stationary) distribution; observe that the value of the second-ranked fitness cannot decrease. A candidate limit is not

hard to come by: the distribution of the random vector $(U, 1, 1, 1, \dots, 1)$ (a $U[0,1]$ variable followed by $N-1$ units) is invariant under the evolution of the Markov chain. We show the following result.

Proposition 1.1 *Let $N \in \mathbb{N}$ and $k \in \{1, 2, \dots, N\}$. If at each step we replace the kth-ranked value by an independent $U[0,1]$ value, then, as $n \to \infty$,*

$$(X_n^{(1)}, X_n^{(2)}, \dots, X_n^{(N)}) \to (0, \dots, 0, U, 1, \dots, 1),$$

in distribution,[10] *where the kth coordinate of the limit vector $U \sim U[0,1]$.*

The process X_n lives on a continuous state space, and it might seem that some fairly sophisticated argument would be needed to show that it has a unique stationary distribution. In fact, we can reduce the problem to a simpler problem on a finite state space as follows.

Sketch of proof of Proposition 1.1 We sketch the argument from [12]. For each $s \in [0,1]$, define the *counting function*[11] $C_n(s) := \sum_{i=1}^{N} \mathbf{1}\{X_n^{(i)} \le s\}$, the number of fitnesses of value at most s at time n. Then $C_n(s)$ is a Markov chain on $\{0, 1, 2, \dots, N\}$. The transition probabilities $p_{x,y} = \mathbb{P}[C_{n+1}(s) = y \mid C_n(s) = x]$ are given for $x \in \{0, \dots, k-1\}$ by $p_{x,x} = 1 - s$ and $p_{x,x+1} = s$, and for $x \in \{k, \dots, N\}$ by $p_{x,x} = s$ and $p_{x,x-1} = 1 - s$. For $s \in (0,1)$, the Markov chain is reducible and all states are transient apart from those in the *recurrent class* $S_k = \{k-1, k\}$. The chain will eventually enter S_k and then never exit. So the problem reduces to that of the two-state restricted chain on S_k. It is easy to compute the stationary distribution and for $s \in (0,1)$, analogously to Theorem 1.7,

$$\lim_{n\to\infty} \mathbb{P}[C_n(s) = x] = \begin{cases} 1-s & \text{if } x = k-1 \\ s & \text{if } x = k \\ 0 & \text{if } n \notin \{k-1, k\} \end{cases}.$$

In particular, for $s \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}[X_n^{(k)} \le s] = \lim_{n\to\infty} \mathbb{P}[C_n(s) \ge k] = s.$$

That is, $X_n^{(k)}$ converges in distribution to a $U[0,1]$ variable. Moreover, if $k > 1$, for any $s \in (0,1)$, $\mathbb{P}[X_n^{(k-1)} \le s] = \mathbb{P}[C_n(s) \ge k-1] \to 1$, which implies that $X_n^{(k-1)}$ converges in probability to 0. Similarly, if $k < N$, for any $s \in (0,1)$, $\mathbb{P}[X_n^{(k+1)} \le s] = \mathbb{P}[C_n(s) \ge k+1] \to 0$, which implies that $X_n^{(k+1)}$ converges in probability to 1. Combining these marginal results, an additional technical step gives the claimed joint convergence: we refer to [12] for details.

1.4 Random Walks

A drunk man will eventually find his way home, but a drunk bird may get lost for ever.

– S. Kakutani's rendering of Pólya's theorem [1, p. 191].

1.4.1 Simple Symmetric Random Walk

The term *random walk* can refer to many different models or classes of models. Although random walks in one dimension had been studied in the context of games of chance, serious study of random

10) That is, for any $x_1, \dots, x_N \in [0,1]$, $\mathbb{P}[X_n^{(1)} \le x_1, \dots, X_n^{(k)} \le x_k, \dots, X_n^{(N)} \le x_N] \to x_k$ if $x_{k+1} \cdots x_N = 1$ and 0 otherwise.

11) "$\mathbf{1}\{\cdot\}$" is the indicator random variable of the appended event: see Section 1.A.

walks as stochastic processes emerged in pioneering works in several branches of science around 1900: Lord Rayleigh's [13] theory of sound developed from about 1880, Bachelier's [14] 1900 model of stock prices, Pearson and Blakeman's [15] 1906 theory of random migration of species, and Einstein's [16] theory of Brownian motion developed during 1905–1908.

In this section, we restrict attention to *simple symmetric random walk* on the integer lattice $\mathbb{Z}^d$. This model had been considered by Lord Rayleigh, but the preeminent early contribution came from George Pólya [17]: we describe his recurrence theorem in the following text. The phrase "random walk" was first applied by statistical pioneer Pearson [18] to a different model in a 1905 letter to *Nature*. We refer to [19] for an overview of a variety of random walk models.

Let $\mathbf{e}_1, \dots, \mathbf{e}_d$ be the standard orthonormal lattice basis vectors for $\mathbb{Z}^d$. Let $X_1, X_2, \dots$ be i.i.d. random vectors with

$$\mathbb{P}[X_1 = \mathbf{e}_i] = \mathbb{P}[X_1 = -\mathbf{e}_i] = \frac{1}{2d}, \text{ for } i \in \{1, \dots, d\}.$$

Let $S_0 = 0$ and $S_n = \sum_{i=1}^n X_i$. Then $(S_n)_{n \in \mathbb{Z}_+}$ is a simple symmetric random walk on $\mathbb{Z}^d$, started from 0; "simple" refers to the fact that the jumps are of size 1.

1.4.2 Pólya's Recurrence Theorem

Clearly (S_n) is a Markov chain; a fundamental question is whether it is recurrent or transient (see Section 1.3.8). Pólya [17] provided the answer in 1921.

Theorem 1.9 (Pólya's theorem) *A simple symmetric random walk on $\mathbb{Z}^d$ is recurrent if $d = 1$ or 2 but transient if $d \geq 3$.*

A basic component in the proof is a combinatorial statement.

Lemma 1.4 *For $d \in \mathbb{N}$ and any $n \in \mathbb{Z}_+$, we have*

$$\mathbb{P}[S_{2n} = 0] = (2d)^{-2n} \binom{2n}{n} \sum_{n_1 + \cdots + n_d = n} \left(\frac{n!}{n_1! \cdots n_d!} \right)^2,$$

where the sum is over d-tuples of nonnegative integers $n_1, \dots, n_d$ that sum to n.

Proof. Each path of length $2n$ (i.e., the possible trajectory for $S_0, S_1, \dots, S_{2n}$) has probability $(2d)^{-2n}$. Any such path that finishes at its starting point must, in each coordinate i, take the same number n_i steps in the positive and negative directions. Enumerating all such paths, we obtain

$$\mathbb{P}[S_{2n} = 0] = (2d)^{-2n} \sum_{n_1 + \cdots + n_d = n} \frac{(2n)!}{(n_1! \cdots n_d!)^2},$$

from which the given formula follows.

Lemma 1.4 and a careful asymptotic analysis using Stirling's formula for $n!$ yields the following result.

Lemma 1.5 *For $d \in \mathbb{N}$, as $n \to \infty$,*

$$n^{d/2} \mathbb{P}[S_{2n} = 0] \to \left(\frac{d}{4\pi} \right)^{d/2}.$$

Proof of Theorem 1.9 Apply the criterion in Lemma 1.3 with Lemma 1.5.

1.4.3 One-dimensional Case; Reflection Principle

We consider in more detail the case $d = 1$. Let $T_a := \min\{n \geq 1 : S_n = a\}$. Theorem 1.9 says that $\mathbb{P}[T_0 < \infty] = 1$. The next result gives the distribution of T_0.

Theorem 1.10 *(i) For any* $n \in \mathbb{Z}_+$, $\mathbb{P}[T_0 = 2n] = (1/(2n-1))\binom{2n}{n}2^{-2n}$.

(ii) $\mathbb{E}[T_0^\alpha] < \infty$ *if and only if* $\alpha < 1/2$.

We proceed by counting sample paths, following the classic treatment by Feller [20, chap. 3]. By an *n-path* we mean a sequence of integers $s_0, \dots, s_n$ where $|s_{i+1} - s_i| = 1$; for an n-path *from a to b* we add the requirement that $s_0 = a$ and $s_n = b$. We view paths as space–time trajectories $(0, s_0), (1, s_1), \dots, (n, s_n)$.

Let $N_n(a, b)$ denote the number of n-paths from a to b. Let $N_n^0(a, b)$ be the number of such paths that visit 0. An n-path from a to b must take $(n + b - a)/2$ positive steps and $(n + a - b)/2$ negative steps, so

$$N_n(a, b) = \binom{n}{\frac{1}{2}(n + b - a)}, \qquad (1.23)$$

where we interpret $\binom{n}{y}$ as 0 if y is not an integer in the range 0 to n.

Lemma 1.6 (Reflection principle) *If $a, b > 0$, then $N_n^0(a, b) = N_n(-a, b)$.*

Proof. Each n-path from $-a$ to b must visit 0 for the first time at some $c \in \{1, \dots, n-1\}$. Reflect in the horizontal (time) axis the segment of this path over $[0, c]$ to obtain an n-path from a to b which visits 0. This reflection is one-to-one.

Theorem 1.11 (Ballot theorem) *If $b > 0$, then the number of n-paths from 0 to b which* do not revisit 0 *is $\frac{b}{n}N_n(0, b)$.*

Proof. The first step of such a path must be 1, so their number is $N_{n-1}(1, b) - N_{n-1}^0(1, b) = N_{n-1}(1, b) - N_{n-1}(-1, b)$, by Lemma 1.6. Now use (1.23).

Theorem 1.12 *If $b \neq 0$ and $n \geq 1$, $\mathbb{P}[T_0 > n, S_n = b] = \frac{|b|}{n}\mathbb{P}[S_n = b]$.*

Proof. Suppose $b > 0$. The event in question occurs if and only if the walk does not visit 0 during $[1, n]$, and $S_n = b$. By the ballot theorem, the number of such paths is $\frac{b}{n}N_n(0, b)$. Similarly for $b < 0$.

At this point, we are ready to prove Theorem 1.10, but first we take a slight detour to illustrate one further variation on "reflection."

Theorem 1.13 *For $a \neq 0$ and $n \geq 1$, $\mathbb{P}[T_a = n] = \frac{|a|}{n}\mathbb{P}[S_n = a]$.*

Proof via time reversal. Fix n. If the trajectory of the original walk up to time n is

$$(S_0, S_1, S_2, \dots, S_n) = \left(0, X_1, X_1 + X_2, \dots, \sum_{i=1}^{n} X_i\right),$$

then the trajectory of the *reversed* walk is

$$(R_0, R_1, R_2, \dots, R_n) = \left(0, X_n, X_n + X_{n-1}, \dots, \sum_{i=1}^{n} X_i\right),$$

that is, the increments are taken in reverse order. The reversed walk has the same distribution as the original walk, because the X_i are i.i.d.

Suppose $a > 0$. The original walk has $S_n = a$ and $T_0 > n$ if and only if the reversed walk has $R_n = a$ and $R_n - R_{n-i} = X_1 + \cdots + X_i > 0$ for all $i \geq 1$, that is, the first visit of the reversed walk to a happens at time n. So $\mathbb{P}[T_a = n] = \mathbb{P}[T_0 > n, S_n = a]$. Now apply Theorem 1.12.

Proof of Theorem 1.10 If $T_0 = 2n$, then $S_{2n-1} = \pm 1$. Thus

$$\mathbb{P}[T_0 = 2n] = \mathbb{P}[T = 2n, S_{2n-1} = 1] + \mathbb{P}[T = 2n, S_{2n-1} = -1]$$

$$= \frac{1}{2}\mathbb{P}[T_0 > 2n-1,\, S_{2n-1} = 1] + \frac{1}{2}\mathbb{P}[T_0 > 2n-1,\, S_{2n-1} = -1].$$

Now by Theorem 1.12,

$$\mathbb{P}[T_0 = 2n] = \frac{1}{2}\cdot\frac{1}{2n-1}\left(\mathbb{P}[S_{2n-1}=1] + \mathbb{P}[S_{2n-1}=-1]\right),$$

and part (i) of the theorem follows from (1.23), after simplification. For part (ii), we have that $\mathbb{E}[T_0^\alpha] = \sum_{n=1}^{\infty}(2n)^\alpha \mathbb{P}[T_0 = 2n]$, and Stirling's formula shows that the summand here is asymptotically a constant times $n^{\alpha-(3/2)}$.

Remark 1.7 *(i) An alternative approach to Theorem 1.10 is via the remarkable identity* $\mathbb{P}[T_0 > 2n] = \mathbb{P}[S_{2n} = 0]$, *which can be verified by a direct but more sophisticated combinatorial argument: see, for example, [21].*

(ii) Yet another approach uses generating functions. For S_n,

$$\psi(s) := \mathbb{E}[s^{S_n}] = \sum_{n=0}^{\infty} s^{2n}\binom{2n}{n}2^{-2n} = \frac{1}{\sqrt{1-s^2}},$$

by (1.23) and then Maclaurin's theorem. Then if ϕ *is the generating function for* T_0, *we can exploit the renewal relation* $\psi(t) = 1 + \phi(t)\psi(t)$ *(see (1.22)) to obtain* $\phi(s) = 1 - \sqrt{1-s^2}$, *from which we can deduce Theorem 1.10 once more.*

1.4.4 Large Deviations and Maxima of Random Walks

In this section, we consider more general one-dimensional random walks in order to illustrate some further concepts. Again we take $S_n = \sum_{i=1}^n X_i$ where the X_i are i.i.d., but now the distribution of X_i will be arbitrary subject to the existence of the mean $\mathbb{E}[X_i] = \mu$. Suppose that $\mu < 0$. The *strong law of large numbers* shows that $n^{-1}S_n \to \mu$, almost surely, as $n \to \infty$. So if $\mu < 0$, then S_n will tend to $-\infty$, and in particular the *maximum* of the walk $M = \max_{n>0} S_n$ is well defined.

There are many applications for the study of M, for example, the modeling of queues (see [22]). What properties does the random variable M possess? We might want to find $\mathbb{P}[M > x]$, for any x, but it is often difficult to obtain exact results; instead we attempt to understand the asymptotic behavior as $x \to \infty$.

Let $\varphi(t) = \mathbb{E}[e^{tX_1}]$ be the moment generating function of the increments. It can be shown that the behavior of $\mathbb{P}[M > x]$ depends on the form of φ. Here we consider only the classical (*light-tailed*) case in which there exists $\gamma > 0$ such that $\varphi(\gamma) = 1$ and $\varphi'(\gamma) < \infty$. For details of the other cases, see [23].

First, Boole's inequality (see Section 1.A) gives

$$\mathbb{P}[M > x] = \mathbb{P}[\cup_{n=1}^{\infty}\{S_n > x\}] \le \sum_{n=1}^{\infty}\mathbb{P}[S_n > x]. \tag{1.24}$$

Now the Chernoff bound (see Section 1.A) implies that, for any $\theta \in [0, \gamma]$,

$$\mathbb{P}[S_n > x] \le e^{-\theta x}\mathbb{E}[e^{\theta S_n}] = e^{-\theta x}(\varphi(\theta))^n;$$

cf. Example 1.4b. We substitute this into (1.24) to obtain

$$\mathbb{P}[M > x] \le e^{-\theta x}\sum_{n=1}^{\infty}(\varphi(\theta))^n = e^{-\theta x}\frac{\varphi(\theta)}{1-\varphi(\theta)},$$

provided $\varphi(\theta) < 1$, which is the case if $\theta \in (0, \gamma)$. For any such θ, we get

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbb{P}[M > x] \leq -\theta - \lim_{x\to\infty} \frac{1}{x} \log(1 - \varphi(\theta)) = -\theta.$$

As $\theta < \gamma$ was arbitrary, we obtain the sharpest bound on letting $\theta \nearrow \gamma$. The matching lower bound can also be proved (see [22]), to conclude that

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}[M > x] = -\gamma.$$

This is an example of a general class of results referred to as *large deviations*: further details of the general theory can be found in [24], for example. These techniques have found use in many application areas including statistical physics: see, for example, [25].

1.5 Markov Chains in Continuous Time

1.5.1 Markov Property, Transition Function, and Chapman–Kolmogorov Relation

In many applications, it is natural to work in *continuous time* rather than the discrete time of Section 1.3. As before, we assume that we have a discrete state space S, but now our Markov chains $X = (X(t))$ have a continuous- time parameter $t \in [0, \infty)$. Continuous time introduces analytical difficulties, which we will not dwell on in this presentation.

As in the discrete-time case, we concentrate on *time-homogeneous* chains, and we will specify the law of $(X(t))$ in line with the Markovian idea that "given the present, the future is independent of the past."

The process $(X(t))$ satisfies the *Markov property* in continuous time if, for all $t, h \geq 0$, all $i, j \in S$, all $0 \leq t_0 < t_1 < \cdots < t_n < t$, and all $i_1, \ldots, i_n \in S$,

$$\mathbb{P}[X(t+h) = j \mid X(t) = i, X(t_n) = i_n, \ldots, X(t_1) = i_1] = p_{ij}(h).$$

Here $p_{ij}(\cdot) = \mathbb{P}[X(t+\cdot) = j \mid X(t) = i]$ is the *transition function* of the Markov chain. As in the discrete-time case, it is convenient to use matrix notation:

$$P(t) = (p_{ij}(t))_{i,j\in S} \text{ given by } p_{ij}(t) = \mathbb{P}_i[X(t) = j],$$

where again a subscript on $\mathbb{P}$ indicates an *initial state*, that is, $\mathbb{P}_i[\cdot] = \mathbb{P}[\cdot \mid X(0) = i]$. We can obtain full information on the law of the Markov chain, analogously to the discrete-time case. For example, for $0 < t_1 < \cdots < t_n$ and $j_1, \ldots, j_n \in S$,

$$\mathbb{P}_i[X(t_1) = j_1, \ldots, X(t_n) = j_n] = p_{ij_1}(t_1) p_{j_1 j_2}(t_2 - t_1) \cdots p_{j_{n-1} j_n}(t_n - t_{n-1}).$$

To this we add information about the initial distribution: for instance, $\mathbb{P}[X(t) = j] = \sum_{i\in S} \mathbb{P}[X(0) = i] p_{ij}(t)$. Here we must have

$$p_{ij}(0) = \delta_{ij} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

We also assume that the transition functions satisfy, for each fixed t, $p_{ij}(t) \geq 0$ for all i, j and $\sum_{j\in S} p_{ij}(t) = 1$ for all i.

To describe our Markov chain now seems a formidable task: we must specify the family of functions $P(t)$. However, we will see in the next section that the Markov property enables a *local* (*infinitesimal*) description. First we state a *global* consequence of the Markov property, namely, Chapman–Kolmogorov relation

$$\text{For any } s, t \geq 0,\ P(s+t) = P(s)P(t). \quad (1.25)$$

The fundamental Markovian relation (1.25) is a special case of the relation

known to probabilists as the *Chapman–Kolmogorov equation*, and which, in its most general form, is often taken as the starting point of the general theory of Markov processes. Physicists refer to a relation such as (1.25) as a *master equation*. The derivation of (1.25) in our setting is direct from the Markov property:

$$\begin{aligned}
&\mathbb{P}_i[X(s+t)=j] \\
&\quad=\sum_{k\in S}\mathbb{P}_i[X(s)=k, X(s+t)=j] \\
&\quad=\sum_{k\in S}\mathbb{P}_i[X(s)=k]\mathbb{P}_i[X(s+t)=j \mid X(s)=k] \\
&\quad=\sum_{k\in S}\mathbb{P}_i[X(s)=k]\mathbb{P}_k[X(t)=j],
\end{aligned}$$

which is the equality for the (i,j) entry in the matrix equation (1.25).

1.5.2 Infinitesimal Rates and *Q*-matrices

In continuous time, there is no smallest time step and so no concept of a one-step transition. Often, however, one can encapsulate the information in the functions $p_{ij}(t)$ in a single fundamental matrix associated with the Markov chain, which will serve as an analogue to the P-matrix in the discrete theory. This is the Q-matrix.

To proceed, we need to assume some regularity. We call the chain *standard* if the transition probabilities are continuous at 0, that is, if $p_{ij}(t) \to p_{ij}(0) = \delta_{ij}$ as $t \downarrow 0$.

Lemma 1.7 *Suppose that X is a standard Markov chain with transition functions $p_{ij}(t)$. Then for each i, j, $p_{ij}(t)$ is a continuous and differentiable function of t. The derivatives $p'_{ij}(t)$ evaluated at $t = 0$ we denote by $q_{ij} := p'_{ij}(0)$; then $0 \le q_{ij} < \infty$ for $i \ne j$ and $0 \le -q_{ii} \le \infty$.*

The proof of this result relies on the Chapman–Kolmogorov relation, but is somewhat involved: see, for example, [26, Section 14.1]. A Taylor's formula expansion now reads

$$\begin{aligned}
p_{ij}(h) &= \mathbb{P}[X(t+h)=j \mid X(t)=i] \\
&= p_{ij}(0) + q_{ij}h + o(h) \qquad (1.26)\\
&= \delta_{ij} + q_{ij}h + o(h), \text{ as } h \downarrow 0. \qquad (1.27)
\end{aligned}$$

So, for $i \ne j$, q_{ij} is the (instantaneous) *transition rate* of the process from state i to state j. It is convenient to define $q_i := -q_{ii}$ for all i (so $q_i \ge 0$). Then q_i is the rate of departure from state i.

We further assume that the chain is *conservative*, meaning

$$\sum_{j\ne i} q_{ij} = q_i < \infty, \text{ for all } i. \qquad (1.28)$$

Note that $\sum_{j\ne i} p_{ij}(t) = 1 - p_{ii}(t)$, so, for example, if S is *finite* we can differentiate to immediately get the equality in (1.28), and then $\sum_{j\ne i} q_{ij} < \infty$ by Lemma 1.7, so a finite Markov chain is always conservative. Note that (1.28) implies that $\sum_j q_{ij} = 0$ and $q_i = \sum_{j\ne i} q_{ij}$, so the rows of Q sum to zero.

The matrix $Q = (q_{ij})_{i,j\in S}$ is called the *transition rate matrix*, the *generator matrix*, or simply the *Q-matrix* of the Markov chain; it effectively describes the chain's *dynamics*. In particular, under reasonable conditions (see the following text) the functions $p_{ij}(\cdot)$ are uniquely determined by Q. Thus, in applications, a Markov process is often *defined* via a Q-matrix and an initial distribution.[12)]

Conversely, given a matrix $Q = (q_{ij})_{i,j\in S}$ with nonpositive diagonal entries and nonnegative entries elsewhere for which (1.28) holds, there always exists a Markov process with Q as transition rate matrix. This fact

12) This is also essentially the approach taken in [6].

can be proved by actually *constructing* the paths of such a process: see Section 1.5.4.

Example 1.18 [*Birth-and-death process*] Here $S = \mathbb{Z}_+$ and $X(t)$ represents a population size at time t. The size of the population increases on a *birth* or decreases on a *death*. The nonzero entries in the Q-matrix are

$$q_{i,i+1} = \lambda_i, \quad i \geq 0 \quad \text{(birth rate in state } i\text{)},$$
$$q_{i,i-1} = \mu_i, \quad i \geq 1 \quad \text{(mortality rate in state } i\text{)},$$
$$q_{0,0} = -\lambda_0 \text{ and } q_{i,i} = -(\lambda_i + \mu_i), \quad i \geq 1.$$

In a *linear* process, $\lambda_i = \lambda i$ and $\mu_i = \mu i$, so λ and μ can be interpreted as *per individual* rates of birth and mortality, respectively.

1.5.3 Kolmogorov Differential Equations

We now consider some differential equations which, given Q, can be used to determine the functions $p_{ij}(\cdot)$. The starting point is the Chapman–Kolmogorov relation $p_{ij}(s+t) = \sum_{k\in S} p_{ik}(s)p_{kj}(t)$. If S is *finite*, say, then it is legitimate to differentiate with respect to s to get

$$p'_{ij}(s+t) = \sum_{k\in S} p'_{ik}(s)p_{kj}(t).$$

Now setting $s = 0$, we obtain

$$p'_{ij}(t) = \sum_{k\in S} q_{ik}p_{kj}(t),$$

which is the *Kolmogorov backward equation*. If instead, we differentiate with respect to t and then put $t = 0$, we obtain (after a change of variable)

$$p'_{ij}(t) = \sum_{k\in S} p_{ik}(t)q_{kj},$$

which is the *Kolmogorov forward equation*. These differential equations are particularly compact in matrix form.

Theorem 1.14 *Given Q satisfying (1.28), we have*

$$P'(t) = P(t)Q,$$
(*Kolmogorov forward equation*);
$$P'(t) = QP(t),$$
(*Kolmogorov backward equation*).

We sketched the derivation in the case where S is finite. In the general case, a proof can be found in, for example, [26, Section 14.2].

Remark 1.8 *Suitable versions of the Kolmogorov differential equations also apply to processes on continuous state spaces, such as* diffusions*, where they take the form of partial differential equations. In this context, the forward equation can be framed as a* Fokker–Planck *equation for the evolution of a probability density. The connections among diffusions, boundary-value problems, and potential theory are explored, for example, in [2]; an approach to Fokker–Planck equations from a more physical perspective can be found in [27].*

We give one example of how to use the Kolmogorov equations, together with generating functions, to compute $P(t)$ from Q.

Example 1.19 [*Homogeneous birth process*] We consider a special case of Example 1.18 with only births, where, for all i, $\lambda_i = \lambda > 0$ and $\mu_i = 0$. So $q_{i,i} = -\lambda$ and $q_{i,i+1} = \lambda$. The Kolmogorov forward equation in this case gives

$$p'_{i,j}(t) = -\lambda p_{i,j}(t) + \lambda p_{i,j-1}(t),$$

where we interpret $p_{i,-1}(t)$ as 0. In particular, if $i = 0$, we have

$$p'_{0,j}(t) = -\lambda p_{0,j}(t) + \lambda p_{0,j-1}(t). \qquad (1.29)$$

The initial conditions are assumed to be $p_{0,0}(0) = 1$ and $p_{0,i}(0) = 0$ for $i \geq 1$, so that the process starts in state 0. Consider the generating function

$$\phi_t(u) = \mathbb{E}[u^{X(t)}] = \sum_{j=0}^{\infty} p_{0,j}(t)u^j, \quad |u| < 1.$$

Multiplying both sides of (1.29) by s^j and summing over j we get

$$\begin{aligned}\frac{\partial}{\partial t}\phi_t(u) &= -\lambda\phi_t(u) + \lambda u\phi_t(u)\\ &= -\lambda(1-u)\phi_t(u).\end{aligned}$$

It follows that $\phi_t(u) = A(u)\mathrm{e}^{-(1-u)\lambda t}$. The initial conditions imply that $\phi_0(u) = 1$, so in fact $A(u) = 1$ here, and $\phi_t(u) = \mathrm{e}^{-(1-u)\lambda t}$, which is the probability generating function of a Poisson distribution with mean λt (see Example 1.2). Hence, $X(t) \sim \mathrm{Po}(\lambda t)$. In fact $X(t)$ is an example of a *Poisson process*: see, for example, [2, 5, 6, 26].

By analogy with the scalar case, under suitable conditions one can define the matrix exponential

$$\exp\{Qt\} = \sum_{k=0}^{\infty} \frac{Q^k t^k}{k!},$$

with $Q^0 = I$ (identity). Then $P(t) = \exp\{Qt\}$ is a formal solution to both the Kolmogorov forward and backward equations. In analytic terminology, Q is the generator of the semigroup P.

1.5.4 Exponential Holding-Time Construction; "Gillespie's Algorithm"

Given the Q-matrix one can construct *sample paths* of a continuous- time Markov chain. The following scheme also tells you how to *simulate* a continuous-time Markov chain.

Suppose the chain starts in a fixed state $X(0) = i$ for $i \in S$. Let $\tau_0 = 0$ and define recursively for $n \geq 0$,

$$\tau_{n+1} = \inf\{t \geq \tau_n : X(t) \neq X(\tau_n)\}.$$

Thus τ_n is the nth *jump time* of X, that is, the nth time at which the process changes its state.

How long does the chain stay in a particular state? We have

$$\begin{aligned}&\mathbb{P}_i[\tau_1 > t + h \mid \tau_1 > t]\\ &\quad = \mathbb{P}_i[\tau_1 > t + h \mid \tau_1 > t, X(t) = i]\\ &\quad = \mathbb{P}_i[\tau_1 > h],\end{aligned}$$

by the Markov property. This *mermoryless* property is indicative of the *exponential distribution* (see Example 1.8a). Recall that $Y \sim \exp(\lambda)$ if $\mathbb{P}[Y > t] = \mathrm{e}^{-\lambda t}$, $t \geq 0$. A calculation shows that

$$\begin{aligned}&\mathbb{P}[Y > t + h \mid Y > t] = \mathbb{P}[Y > h]\\ &\quad = \mathrm{e}^{-\lambda h} = 1 - \lambda h + o(h), \qquad (1.30)\end{aligned}$$

as $h \to 0$.

In fact, the exponential distribution is essentially the only distribution with this property. So it turns out that τ_1 is exponential. A heuristic calculation, which can be justified, suggests that $\mathbb{P}_i[\tau_1 > h] \sim \mathbb{P}_i[X(h) = i] = p_{ii}(h) = 1 - q_i h + o(h)$. A comparison with (1.30) suggests that $\tau_1 \sim \exp(q_i)$.

When the chain does jump, where does it go? Now, for $j \neq i$, $\mathbb{P}[X(t + h) = j$

$| X(t) = i] = p_{ij}(h) = q_{ij}h + o(h)$, while $\mathbb{P}[X(t+h) \neq i \mid X(t) = i] = q_i h + o(h)$, so a conditional probability calculations gives

$$\mathbb{P}[X(t+h) = j \mid X(t) = i, X(t+h) \neq i] = \frac{q_{ij}}{q_i} + o(1).$$

Careful argument along these lines (see, e.g., [26, Section 14.3]) gives the next result.[13)]

Theorem 1.15 *Under the law $\mathbb{P}_i$ of the Markov chain started in $X(0) = i$, the random variables τ_1 and $X(\tau_1)$ are independent. The distribution of τ_1 is exponential with rate q_i. Moreover, $\mathbb{P}_i[X(\tau_1) = j] = q_{ij}/q_i$.*

Perhaps the most striking aspect of this result is that the holding time and the jump destination are independent. Theorem 1.15 tells us how to construct the Markov chain, by iterating the following procedure.

- Given τ_n and $X(\tau_n) = i$, generate an $\exp(q_i)$ random variable Y_n (this is easily done via $Y_n = -q_i^{-1} \log U_n$, where $U_n \sim U[0,1]$). Set $\tau_{n+1} = \tau_n + Y_n$.
- Select the next state $X(\tau_{n+1})$ according to the distribution q_{ij}/q_i.

Although this standard construction of Markov chain sample paths goes back to classical work of Doeblin, Doob, Feller, and others in the 1940s, and was even implemented by Kendall and Bartlett in pioneering computer simulations in the early 1950s, the scheme is known in certain applied circles as "*Gillespie's algorithm*" after Gillespie's 1977 paper that rederived the construction in the context of chemical reaction modeling.

13) Actually Theorem 1.15 assumes that we are working with the *minimal* version of the process: see, for example, [6, 26] for details of this technical point.

Remark 1.9 *Let $X_n^* = X(\tau_n)$. Then $(X_n^*)_{n \in \mathbb{Z}_+}$ defines a discrete-time Markov chain, called the* jump chain *associated with $X(t)$, with one-step transitions*

$$p_{ij}^* = \begin{cases} \frac{q_{ij}}{q_i} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases},$$

as long as $q_i > 0$. If $q_i = 0$, then $p_{ii}^* = 1$, that is, i is an absorbing state.

1.5.5 Resolvent Computations

Consider a Markov chain with transition functions $p_{ij}(t)$ determined by its generator matrix Q. The Laplace transform of p_{ij} is r_{ij} given by

$$r_{ij}(\lambda) = \int_0^\infty \mathrm{e}^{-\lambda t} p_{ij}(t) \mathrm{d}t. \tag{1.31}$$

Then $R(\lambda) = (r_{ij}(\lambda))_{i,j \in S}$ is the *resolvent matrix* of the chain. A formal calculation, which can be justified under the conditions in force in this section, shows that $R(\lambda)$ can be expressed as the matrix inverse $R(\lambda) = (\lambda I - Q)^{-1}$, where I is the identity. See Chapter 15 of the present volume for background on Laplace transforms.

Let τ be an $\exp(\lambda)$ random variable, independent of the Markov chain. Then

$$\begin{aligned} \lambda r_{ij}(\lambda) &= \int_0^\infty \lambda \mathrm{e}^{-\lambda t} p_{ij}(t) \mathrm{d}t \\ &= \int_0^\infty \mathbb{P}[\tau \in \mathrm{d}t] \mathbb{P}_i[X(t) = j \mid \tau = t], \end{aligned}$$

which is just $\mathbb{P}_i[X(\tau) = j]$, the probability that, starting from state i, the chain is in state j at the random time τ.

Resolvents play an important role in the theoretical development of Markov processes, and in particular in the abstract semigroup approach to the theory. Here, however, we view the resolvent as a computational tool, which enables, in principle,

calculation of probabilities and hitting-time distributions.

Specifically, (1.31) implies that $p_{ij}(t)$ can be recovered by inverting its Laplace transform $\hat{p}_{ij}(\lambda) = r_{ij}(\lambda)$. Moreover, let $T_i := \inf\{t \geq 0 : X(t) = i\}$, the first hitting time of state i. Write $F_{ij}(t) := \mathbb{P}_i[T_j \leq t]$, and set $f_{ij}(t) = F'_{ij}(t)$ for the density of the hitting-time distribution. We proceed analogously to the discrete argument for the proof of Lemma 1.3. An application of the *strong Markov property* for continuous-time chains (cf Section 1.3.4) gives,

$$p_{ij}(t) = \int_0^t f_{ij}(s)p_{jj}(t-s)\mathrm{d}s.$$

The convolution theorem for Laplace transforms (see Chapter 15) implies that the Laplace transform of f_{ij} is given by

$$\hat{f}_{ij}(\lambda) = \frac{r_{ij}(\lambda)}{r_{jj}(\lambda)}. \tag{1.32}$$

In the next section, we give some examples of using resolvent ideas in computations.

1.5.6 Example: A Model of Deposition, Diffusion, and Adsorption

We describe a continuous-time Markov model of deposition of particles that subsequently perform random walks and interact to form barriers according to an occupation criterion, inspired by models of submonolayer film growth [28, 29]. Particles arrive randomly one by one on a one-dimensional substrate $S_N := \{0, 1, \dots, N+1\}$ and diffuse until $M \geq 2$ particles end up at the same site, when they clump together ("nucleate") to form an "island." Islands form absorbing barriers with respect to the diffusion of other particles. We assume that initially, sites 0 and $N+1$ are occupied by M particles (so are already islands) but all other sites are empty.

The Markov dynamics are as follows.

- At each site $x \in S_N$, new particles arrive independently at rate $\rho > 0$.
- If at any time a site is occupied by M or more particles, all those particles are held in place and are *inactive*. Particles that are not inactive are *active*.
- Each active particle independently performs a symmetric simple random walk at rate 1, that is, from x it jumps to $x+1$ or $x-1$ each at rate $1/2$.

A state ω of the Markov process is a vector of the occupancies of the sites $1, \dots, N$ (it is not necessary to keep track of the occupancies of 0 or $N+1$): $\omega(x)$ is the number of particles at site x. We can simulate the process via the exponential holding-time construction of Section 1.5.4. To do so, we need to keep track of $T(\omega) = \sum_{1 \leq x \leq N} \omega(x)\mathbf{1}\{\omega(x) < M\}$, the total number of active particles in state ω. The waiting time in a state ω is then exponential with parameter $T(\omega) + N\rho$; at the end of this time, with probability $T(\omega)/(T(\omega) + N\rho)$, one of the active particles jumps (chosen uniformly from all active particles, and equally likely to be a jump left or right), else a new particle arrives at a uniform random site in $\{1, \dots, N\}$.

An analysis of the general model just described would be interesting but is beyond the scope of this presentation. We use small examples (in terms of M and N) to illustrate the resolvent methods described in Section 1.5.5. For simplicity, we take $M = 2$ and stop the process the first time that two particles occupy any internal site. Configurations can be viewed as elements of $\{0, 1, 2\}^{\{1,2,\dots,N\}}$, but symmetry can be used to further reduce the state space for our questions of interest.

1.5.6.1 $N = 1$

Take $N = 1$. The state space for our Markov chain $X(t)$ is $\{0, 1, 2\}$, the number of particles in position 1, with $X(0) = 0$, and 2 as the absorbing state. Clearly $X(t) = 2$ eventually; the only question is how long we have to wait for absorption. The answer is not trivial, even in this minimal example.

The generator matrix for the Markov chain is

$$Q = \begin{pmatrix} -\rho & \rho & 0 \\ 1 & -1-\rho & \rho \\ 0 & 0 & 0 \end{pmatrix}, \text{ and so}$$

$$\lambda I - Q = \begin{pmatrix} \lambda+\rho & -\rho & 0 \\ -1 & 1+\lambda+\rho & -\rho \\ 0 & 0 & \lambda \end{pmatrix}.$$

To work out $p_{02}(t)$, we compute $r_{02}(\lambda)$:

$$r_{02}(\lambda) = (\lambda I - Q)^{-1}_{02} = \frac{\det \begin{pmatrix} -\rho & 0 \\ 1+\lambda+\rho & -\rho \end{pmatrix}}{\det(\lambda I - Q)} = \frac{\rho^2}{\lambda((\lambda+\rho)^2+\lambda)}.$$

Inverting the Laplace transform, we obtain

$$1 - p_{02}(t) = e^{-((1+2\rho)/(2))t} \left(\cosh\left(\left(\frac{t}{2}\right)\sqrt{1+4\rho}\right) + \frac{1+2\rho}{\sqrt{1+4\rho}} \sinh\left(\left(\frac{t}{2}\right)\sqrt{1+4\rho}\right) \right)$$

$$\sim \frac{1}{2}\left(1 + \frac{1+2\rho}{\sqrt{1+4\rho}}\right) \exp\left\{-\frac{t}{2}\left(1+2\rho-\sqrt{1+4\rho}\right)\right\}, \quad \text{as } t \to \infty.$$

For example, if $\rho = 3/4$, the exact expression simplifies to

$$p_{02}(t) = 1 - \frac{9}{8}e^{-t/4} + \frac{1}{8}e^{-9t/4}.$$

As in this case, 2 is absorbing, $\mathbb{P}[T_2 \le t] = \mathbb{P}[X(t) = 2] = p_{02}(t)$, so

$$f_{02}(t) = \frac{d}{dt}p_{02}(t) = \frac{9}{32}(e^{-t/4} - e^{-9t/4}),$$

in the $\rho = 3/4$ example. The same answer can be obtained using (1.32).

1.5.6.2 $N = 3$

For the problem of the distribution of the point of the first collision, the first nontrivial case is $N = 3$. Making use of the symmetry, we can now describe the Markov chain with the 8 states

$$(0,0,0),\ (1,0,0),\ (0,1,0),\ (1,1,0),\ (1,0,1),\ (1,1,1),\ (2,*,*),\ (*,2,*),$$

in that order, where each $*$ indicates either a 0 or a 1; so for example $(1,0,0)$ stands for $(1,0,0)$ or $(0,0,1)$. The generator matrix is now

$$Q = \begin{pmatrix} -3\rho & 2\rho & \rho & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & -1-3\rho & \frac{1}{2} & \rho & \rho & 0 & \rho & 0 \\ 0 & 1 & -1-3\rho & 2\rho & 0 & 0 & 0 & \rho \\ 0 & 0 & \frac{1}{2} & -2-3\rho & \frac{1}{2} & \rho & \frac{1}{2}+\rho & \frac{1}{2}+\rho \\ 0 & 1 & 0 & 1 & -2-3\rho & \rho & 2\rho & 0 \\ 0 & 0 & 0 & 1 & 0 & -3-3\rho & 1+2\rho & 1+\rho \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The probability of the first collision occurring at the midpoint is

$$z(\rho) = \mathbb{P}\big[\lim_{t\to\infty} X(t) = (*, 2, *)\big].$$

According to MAPLE, inverting the appropriate Laplace transform gives

$$z(\rho) = \frac{1}{9} \cdot \frac{486\rho^4 + 1354\rho^3 + 1375\rho^2 + 598\rho + 87}{162\rho^4 + 414\rho^3 + 392\rho^2 + 160\rho + 23}.$$

In particular, if deposition dominates diffusion,

$$\lim_{\rho\to\infty} z(\rho) = \frac{1}{9} \cdot \frac{486}{162} = \frac{1}{3},$$

which is as it should be!

1.6 Gibbs and Markov Random Fields

We have so far focused on stochastic processes that vary through time. In this section and in Section 1.7, we take a detour into *spatial processes*. The role of space in our models is played by an underlying *graph* structure, describing vertices V and the edges E that connect them. This section is devoted to *random fields*, that is, ensembles of random variables associated with the vertices subject to certain constraints imposed by the edges.

Let $G = (V, E)$ be an undirected finite graph with vertex set V and edge set E consisting of pairs (i, j) where $i, j \in V$; $(i, j) \in E$ indicates an edge between vertices i and j.[14] With this graph, we associate random variables $\{X_i\}$ for $i \in V$; to keep things simple, we take $X_i \in \{-1, 1\}$.

We consider two ways of specifying these random variables – as Gibbs or Markov random fields – and the relationship between the two. Finally, we will consider how to use ideas from Markov chains (Section 1.3) to simulate a Markov random field.

14) Our graphs are undirected, which means that $(i, j) = (j, i)$ is an unordered pair.

1.6.1 Gibbs Random Field

To define a Gibbs random field, we need the concept of a *clique*. A clique in $G = (V, E)$ is a subset K of V such that E contains all the possible edges between members of K; we include the set of no vertices $\emptyset$ as a clique. Let $\mathcal{K}$ be the set of all cliques of graph G.

Let $\mathbf{x} = (x_1, \ldots, x_{|V|}) \in \{-1, 1\}^V$ denote an assignment of a value ± 1 to each vertex of G. For $K \subseteq V$ we write $\mathbf{x}|_K$ for the restriction of $\mathbf{x}$ to K. The random variables $\{X_i\}_{i\in V}$ on G constitute a *Gibbs random field* if, for all $\mathbf{x} \in \{-1, 1\}^V$,

$$\mathbb{P}[X_i = x_i \text{ for all } i \in V] = \frac{1}{Z} \exp\Big\{ \sum_{K\in\mathcal{K}} f_K(\mathbf{x}|_K) \Big\}, \tag{1.33}$$

where $f_K : \{-1, 1\}^{|K|} \to \mathbb{R}$ for each clique K, and Z is a normalization:

$$Z = \sum_{\mathbf{x}\in\{-1,1\}^V} \exp\Big\{ \sum_{K\in\mathcal{K}} f_K(\mathbf{x}|_K) \Big\}.$$

To see why this is a natural definition for the law of $\{X_i\}$ in many situations, consider associating an *energy* function $\mathcal{E} : \{-1, 1\}^V \to \mathbb{R}$ to the states. We seek the maximum entropy distribution for $\{X_i\}_{i\in V}$ for a given mean energy. So we want to find the probability mass function, f, on $\{-1, 1\}^V$ that maximizes $-\sum_{\mathbf{x}\in\{-1,1\}^V} f(\mathbf{x}) \log f(\mathbf{x})$ subject to $\sum_{\mathbf{x}\in\{-1,1\}^V} f(\mathbf{x})\mathcal{E}(x) = \text{const}$. One can show this is achieved by

$$f(\mathbf{x}) \propto \exp\{-\beta\mathcal{E}(x)\},$$

where $\beta \in (0, \infty)$ is chosen to obtain the required energy. (The analogy here is between β and the inverse temperature $1/(kT)$ in thermodynamics.) Now if $\mathcal{E}$ can be decomposed as a sum over the cliques of the graph, we recover (1.33).

Example 1.20 [Ising model] Consider the $N \times N$ grid in two dimensions. We label the vertices by members of the set $L_N = \{1, \dots, N\} \times \{1, \dots, N\}$. We put an edge between $i = (i_1, i_2)$ and $j = (j_1, j_2)$ if $|i_1 - j_1| + |i_2 - j_2| = 1$; this adjacency condition we write as $i \sim j$. Here the clique set $\mathcal{K}$ is made up of the empty set, singleton nodes, and pairs of nodes that are distance one apart in the lattice.

Given constants $\beta > 0$, $J > 0$, and $h \in \mathbb{R}$ we consider the Gibbs random field with probability mass function $f(\mathbf{x}) = Z^{-1} \exp\{-\beta \mathcal{E}(x)\}$, where

$$\mathcal{E}(x) = -J \sum_{i,j \in L_N : i \sim j} X_i X_j - h \sum_{i \in L_N} X_i.$$

The sum over pairs of nodes (the *interaction* term) means that neighboring nodes have a propensity to be in the same state. The second (*external field*) term leads to nodes more likely to be in either of the states 1 or -1, depending on the sign of h.

This model has been studied widely as a model for ferromagnetism and was initially proposed by Ising [30] under the guidance of Lenz. The *Potts model* is a generalization with q-valued states and more general interactions: see [31].

1.6.2 Markov Random Field

We now consider a second specification of random field that adapts the Markov property to spatial processes. For a given subset $W \subseteq V$, we define its boundary as

$$\partial W = \{v \in V \setminus W : (v, w) \in E \text{ for some } w \in W\}.$$

The concept of Markov random field extends the temporal Markov property (1.12), which said that, conditional on the previous states of the process, the future depends on the past only through the present, to a spatial (or topological) one. This "Markov property" will say that the state of nodes in some set of vertices W conditioned on the state of all the other vertices only depends on the state of the vertices in ∂W.

The random variables $\{X_i\}_{i \in V}$ on G constitute a *Markov random field* if

- they have a *positive* probability mass function,

$$\mathbb{P}[\{X_i\}_{i \in V} = \mathbf{x}] > 0, \quad \text{for all } \mathbf{x} \in \{-1, 1\}^V,$$

- and obey the global *Markov property*: for all $W \subseteq V$,

$$\mathbb{P}[\{X_i\}_{i \in W} = \mathbf{x}|_W \mid \{X_i\}_{i \in V \setminus W} = \mathbf{x}|_{V \setminus W}] = \mathbb{P}[\{X_i\}_{i \in W} = \mathbf{x}|_W \mid \{X_i\}_{i \in \partial W} = \mathbf{x}|_{\partial W}].$$

Example 1.21 As in Example 1.20, consider a random field taking values ± 1 on the vertices of the $N \times N$ lattice. We specify the (conditional) probability that a vertex, i, is in state 1, given the states of its neighbors to be

$$\frac{e^{\beta(h + J y_i)}}{e^{-\beta(h + J y_i)} + e^{\beta(h + J y_i)}}, \quad \text{where } y_i = \sum_{j \sim i} x_j. \tag{1.34}$$

Here $\beta > 0$, $J > 0$, and $h \in \mathbb{R}$ are parameters. The larger y_i, which is the number of neighbours of i with spin $+1$ minus the

number with spin −1, the greater the probability that vertex i will itself have state 1.

1.6.3 Connection Between Gibbs and Markov Random Fields

In Examples 1.20 and 1.21, we have used the same notation for the parameters. In fact, both specifications (one Gibbs, the other Markov) define the same probability measure on $\{-1, 1\}^V$. This is an example of the following result.

Theorem 1.16 (Hammersley–Clifford theorem) *The ensemble of random variables $\{X_i\}_{i \in V}$ on G is a Markov random field if and only if it is a Gibbs random field with a positive probability mass function.*

A proof can be found in [31]. From this point forward, we will use the terms Gibbs random field and Markov random field interchangeably.

1.6.4 Simulation Using Markov Chain Monte Carlo

Direct simulation of a Gibbs random field on a graph is computationally difficult because the calculation of the normalizing constant, Z, requires a sum over all the possible configurations. In many situations, this is impractical. Here we consider an alternative way to simulate a Gibbs random field making use of Markov chains.

We saw in Section 1.3.7 that an irreducible, aperiodic Markov chain converges to its stationary distribution. The idea now is to design a Markov chain on the state space $\{-1, 1\}^V$ whose stationary distribution coincides with the desired Gibbs random field. We simulate the Markov chain for a long time to obtain what should be a distribution close to stationarity, and hence, a good approximation to a realization of the Gibbs random field.

We initialize the Markov chain with any initial state $\sigma \in \{-1, 1\}^V$. To update the state of the chain, we randomly select a vertex uniformly from all vertices in the graph. We will update the state associated with this vertex by randomly selecting a new state using the conditional probabilities given, subject to the neighboring vertices' states, taking advantage of the Markov random field description. For example, in Example 1.21, we set the node state to 1 with probability given by (1.34) and to −1, otherwise.

It is easy to check that this Markov chain is irreducible, aperiodic, and has the required stationary distribution. This is an example of a more general methodology of using a Markov chain with simple update steps to simulate from a distribution that is computationally difficult to evaluate directly, called *Markov chain Monte Carlo*. Specifically, we have used a *Gibbs sampler* here, but there are many other schemes for creating a Markov chain with the correct stationary distribution. Many of these techniques have been developed within the setting of Bayesian statistics but have applications in many other fields, including spin glass models and theoretical chemistry.

1.7 Percolation

Consider the infinite square lattice $\mathbb{Z}^2$. Independently, for each edge in the lattice, the edge is declared *open* with probability p; else (with probability $1 - p$) it is *closed*. This model is called *bond percolation* on the square lattice.

For two vertices $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^2$ write $\mathbf{x} \longleftrightarrow \mathbf{y}$ if $\mathbf{x}$ and $\mathbf{y}$ are joined by a path consisting of open edges in the percolation model on $\mathbb{Z}^2$.

The *open cluster* containing vertex $\mathbf{x}$, $C(\mathbf{x})$, is the (random) set of all vertices joined to $\mathbf{x}$ by an open path

$$C(\mathbf{x}) := \{\mathbf{y} \in \mathbb{Z}^2 : \mathbf{y} \longleftrightarrow \mathbf{x}\}.$$

A fundamental question in percolation regards the nature of $C(\mathbf{0})$, the open cluster at $\mathbf{0}$, and how its (statistical) properties depend on the parameter p.

We write $\mathbb{P}_p$ for the probability associated with percolation with parameter p. Write $|C(\mathbf{0})|$ for the number of vertices in the open cluster at $\mathbf{0}$. The *percolation probability* is

$$\theta(p) = \mathbb{P}_p[|C(\mathbf{0})| = \infty].$$

By translation invariance, $\theta(p) \in [0, 1]$ is not specific to the origin: $\mathbb{P}_p[|C(\mathbf{x})| = \infty] = \theta(p)$ for any $\mathbf{x}$.

Let H_∞ be the event that $|C(\mathbf{x})| = \infty$ for *some* $\mathbf{x}$. It is not hard to show that

$$\theta(p) = 0 \implies \mathbb{P}_p[H_\infty] = 0$$
$$\theta(p) > 0 \implies \mathbb{P}_p[H_\infty] = 1.$$

We state a fundamental result that may seem obvious; the proof we give, due to Hammersley, demonstrates the effectiveness of another probabilistic tool: *coupling*.

Lemma 1.8 *$\theta(p)$ is nondecreasing as a function of p.*

Proof. List the edges of the lattice $\mathbb{Z}^2$ in some order as $e_1, e_2, \ldots$. Let $U_1, U_2, \ldots$ be independent uniform random variables on $[0, 1]$. Assign U_i to edge e_i.

Let $E_p = \{e_i : U_i \le p\}$. Then E_p is the set of open edges in bond percolation with parameter p. This construction *couples* bond percolation models for every $p \in [0, 1]$ in a monotone way: if $e_i \in E_p$ then $e_i \in E_q$ for all $q \ge p$.

Let $C_p(\mathbf{0})$ denote the cluster containing $\mathbf{0}$ using edges in E_p. If $p \le q$, then by construction $C_p(\mathbf{0}) \subseteq C_q(\mathbf{0})$. So $\{|C_p(\mathbf{0})| = \infty\} \subseteq \{|C_q(\mathbf{0})| = \infty\}$, and hence, $\theta(p) \le \theta(q)$.

As $\theta(p)$ is nondecreasing, and clearly $\theta(0) = 0$ and $\theta(1) = 1$, there must be some threshold value

$$p_c := \inf\{p \in [0, 1] : \theta(p) > 0\}.$$

So for $p < p_c$, $\mathbb{P}_p[H_\infty] = 0$, while for $p > p_c$, $\mathbb{P}_p[H_\infty] = 1$.

The first question is this: is there a nontrivial *phase transition*, that is, is $0 < p_c < 1$? This question was answered by Broadbent and Hammersley in the late 1950s.

Proposition 1.2 $1/3 \le p_c \le 2/3$.

Proof of $p_c \ge 1/3$ Let A_n be the event that there exists a self-avoiding open path starting at $\mathbf{0}$ of length n. Then

$$A_1 \supseteq A_2 \supseteq A_3 \cdots \quad \text{and}$$
$$\bigcap_{n=1}^{\infty} A_n = \{|C(\mathbf{0})| = \infty\}.$$

So $\theta(p) = \mathbb{P}_p[|C(\mathbf{0})| = \infty] = \lim_{n\to\infty} \mathbb{P}_p[A_n]$, by continuity of probability measure (see Section 1.A). Let Γ_n be the set of all possible self-avoiding paths of length n starting at $\mathbf{0}$. Then

$$\begin{aligned}\mathbb{P}_p[A_n] &= \mathbb{P}_p \bigcup_{\gamma\in\Gamma_n} \{\gamma \text{ is open}\} \\ &\le \sum_{\gamma\in\Gamma_n} \mathbb{P}_p[\gamma \text{ is open}] = |\Gamma_n| p^n \\ &\le 4 \cdot 3^{n-1} \cdot p^n,\end{aligned}$$

which tends to 0 if $p < 1/3$. So $p_c \ge 1/3$.

On the basis of pioneering Monte Carlo simulations, Hammersley conjectured that

p_c was 1/2. Harris proved in 1960 that $\theta(1/2) = 0$, which implies that $p_c \geq 1/2$. It was not until 1980 that a seminal paper of Kesten settled things.

Theorem 1.17 (Harris 1960, Kesten 1980) $p_c = 1/2$.

Harris's result $\theta(1/2) = 0$ thus means that $\theta(p_c) = 0$; this is conjectured to be the case in many percolation models (e.g., on $\mathbb{Z}^d$, it is proved for $d = 2$ and $d \geq 19$, but is conjectured to hold for all $d \geq 2$). In recent years, there has been much interest in the detailed structure of percolation when $p = p_c$. The *Schramm–Loewner evolution* has provided an important new mathematical tool to investigate physical predictions, which often originated in conformal field theory; see [32].

1.8 Further Reading

A wealth of information on stochastic processes and the tools that we have introduced here can be found in [2, 5, 9, 20, 26], for example. All of those books cover Markov chains. The general theory of Markov processes can be found in [2, 9], which also cover the connection to semigroup theory. Feller gives a masterly presentation of random walks and generating functions [20] and Laplace transforms, characteristic functions, and their applications [9]. Branching processes can be found in [5, 20]. A thorough treatment of percolation is presented in [33]. We have said almost nothing here about Brownian motion or diffusions, for which we refer the reader to Chapter 3 of this volume as well as [2, 5, 9, 26]. Physicists and mathematicians alike find it hard not to be struck by the beauty of the connection between random walks and electrical networks, as exposited in [34]. Applications of stochastic processes in physics and related fields are specifically treated in [27, 35]; the array of applications of random walks alone is indicated in [19, 36, 37].

1.A Appendix: Some Results from Probability Theory

There is no other simple mathematical theory that is so badly taught to physicists as probability.

– R. F. Streater [38, p. 19].

Essentially, the theory of probability is nothing but good common sense reduced to mathematics.

– P.-S. de Laplace, Essai philosophique sur les probabilités, 1813.

Kolmogorov's 1933 axiomatization of probability on the mathematical foundation of measure theory was fundamental to the development of the subject and is essential for understanding the modern theory. Many excellent textbook treatments are available. Here we emphasize a few points directly relevant for the rest of this chapter.

1.A.1 Set Theory Notation

A *set* is a collection of *elements*. The set of no elements is the *empty set* $\emptyset$. Finite nonempty sets can be listed as $S = \{a_1, \dots, a_n\}$. If a set S contains an element a, we write $a \in S$. A set R is a *subset* of a set S, written $R \subseteq S$, if every $a \in R$ also satisfies $a \in S$. For two sets S and T, their *intersection* is $S \cap T$, the set of elements that are in *both* A and B, and their *union* is

$S \cup T$, the set of elements in at least one of S or T. For two sets S and T, "S minus T" is the set $S \setminus T = \{a \in S : a \notin T\}$, the set of elements that are in S but not in T.

Note that $S \cap \emptyset = \emptyset$, $S \cup \emptyset = S$, and $S \setminus \emptyset = S$.

1.A.2 Probability Spaces

Suppose we perform an experiment that gives a random outcome. Let Ω denote the set of all possible outcomes: the *sample space*. To start with, we take Ω to be *discrete*, which means it is *finite* or *countably infinite*. This means that we can write Ω as a (possibly infinite) list:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\},$$

where $\omega_1, \omega_2, \omega_3, \dots$ are the possible outcomes to our experiment.

A set $A \subseteq \Omega$ is called an *event*. Given events $A, B \subseteq \Omega$, we can build new events using the operations of set theory:

- $A \cup B$ ("A or B"), the event that A happens, or B happens, or both.
- $A \cap B$ ("A and B"), the event that A and B both happen.

Two events A and B are called *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$.

We want to assign probabilities to events. Let Ω be a nonempty discrete sample space. A function $\mathbb{P}$ that gives a value $\mathbb{P}[A] \in [0, 1]$ for every subset $A \subseteq \Omega$ is called a discrete *probability measure* on Ω if

(P1) $\mathbb{P}[\emptyset] = 0$ and $\mathbb{P}[\Omega] = 1$;

(P2) For any $A_1, A_2, \dots$, pairwise disjoint subsets of Ω (so $A_i \cap A_j = \emptyset$ for $i \neq j$),

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i] \quad (\sigma\text{-additivity}).$$

Given Ω and a probability measure $\mathbb{P}$, we call $(\Omega, \mathbb{P})$ a *discrete probability space*.

Remark 1.10 *For nondiscrete sample spaces, we may not be able to assign probabilities to* all *subsets of Ω in a sensible way, and so smaller collections of events are required. In this appendix, we treat the discrete case only, as is sufficient for most (but not all) of the discussion in this chapter. For the more general case, which requires a deeper understanding of measure theory, there are many excellent treatments, such as [1, 2, 39, 40].*

For an event $A \subseteq \Omega$, we define its *complement*, denoted A^c and read "not A", to be $A^c := \Omega \setminus A = \{\omega \in \Omega : \omega \notin A\}$. Note that $(A^c)^c = A$, $A \cap A^c = \emptyset$, and $A \cup A^c = \Omega$.

If $(\Omega, \mathbb{P})$ is a discrete probability space, then

- For $A \subseteq \Omega$, $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$;
- If $A, B \subseteq \Omega$ and $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$ (monotonicity);
- If $A, B \subseteq \Omega$, then $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. In particular, if $A \cap B = \emptyset$, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$.

It follows from this last statement that $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$; more generally, we have the elementary but useful *Boole's inequality*: $\mathbb{P}[\cup_n A_n] \leq \sum_n \mathbb{P}[A_n]$.

At several points we use the *continuity property of probability measures*:

- If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ are events, then $\mathbb{P}[\cup_{i=1}^{\infty} A_i] = \lim_{n\to\infty} \mathbb{P}[A_n]$.
- If $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ are events, then $\mathbb{P}[\cap_{i=1}^{\infty} A_i] = \lim_{n\to\infty} \mathbb{P}[A_n]$.

To see the first statement, we can write $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} (A_i \setminus A_{i-1})$, where we set $A_0 =$

$\emptyset$, and the latter union is over pairwise disjoint events. So, by σ-additivity,

$$\mathbb{P}[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} \mathbb{P}[A_i \setminus A_{i-1}] = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}[A_i \setminus A_{i-1}].$$

But $\sum_{i=1}^{n} \mathbb{P}[A_i \setminus A_{i-1}] = \mathbb{P}[\cup_{i=1}^{n}(A_i \setminus A_{i-1})] = \mathbb{P}[A_n]$, giving the result. The second statement is analogous.

1.A.3 Conditional Probability and Independence of Events

If A and B are events with $\mathbb{P}[B] > 0$ then the *conditional probability* $\mathbb{P}[A \mid B]$ of A given B is defined by

$$\mathbb{P}[A \mid B] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

A countable collection of events $E_1, E_2, \ldots$ is called a *partition* of Ω if

(a) for all i, $E_i \subseteq \Omega$ and $E_i \neq \emptyset$;

(b) for $i \neq j$, $E_i \cap E_j = \emptyset$ (the events are disjoint);

(c) $\bigcup_i E_i = \Omega$ (the events fill the sample space).

Let $E_1, E_2, \ldots$ be a partition of Ω. Following from the definitions is the basic *law of total probability*, which states that for all $A \subseteq \Omega$,

$$\mathbb{P}[A] = \sum_i \mathbb{P}[E_i]\mathbb{P}[A \mid E_i].$$

A countable collection $(A_i, i \in I)$ of events is called *independent* if, for every *finite* subset $J \subseteq I$,

$$\mathbb{P}\Big[\bigcap_{j\in J} A_j\Big] = \prod_{j\in J} \mathbb{P}[A_j].$$

In particular, two events A and B are independent if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$ (i.e., if $\mathbb{P}[B] > 0$, $\mathbb{P}[A \mid B] = \mathbb{P}[A]$).

1.A.4 Random Variables and Expectation

Let $(\Omega, \mathbb{P})$ be a discrete probability space. A function $X : \Omega \to \mathbb{R}$ is a *random variable*. So each $\omega \in \Omega$ is mapped to a real number $X(\omega)$. The set of possible values for X is $X(\Omega) = \{X(\omega) : \omega \in \Omega\} \subset \mathbb{R}$. Notice that because Ω is discrete, $X(\Omega)$ must be also.

If X and Y are two random variables on $(\Omega, \mathbb{P})$, then $X + Y$, XY, and so on, are also random variables. For example, $(X + Y)(\omega) = X(\omega) + Y(\omega)$. In some cases (such as the expected hitting times defined at (1.18)), we extend the domain of a random variable from $\mathbb{R}$ to $\mathbb{R} \cup \{\infty\}$.

The *probability mass function* of a discrete random variable X is the collection $\mathbb{P}[X = x]$ for all $x \in X(\Omega)$. The *distribution function* of X is $F : \mathbb{R} \to [0, 1]$ given by $F(x) = \mathbb{P}[X \leq x]$.

Example 1.22 [Bernoulli and binomial distributions] Let n be a positive integer and $p \in [0, 1]$. We say X has a *binomial distribution* with parameters (n, p), written $X \sim \text{Bin}(n, p)$, if $\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$ for $k \in \{0, 1, \ldots, n\}$. The case $Y \sim \text{Bin}(1, p)$ is the *Bernoulli* distribution; here $\mathbb{P}[Y = 1] = p = 1 - \mathbb{P}[Y = 0]$ and we write $Y \sim \text{Be}(p)$. The binomial distribution has the following interpretation: Perform n independent "trials" (e.g., coin tosses) each with probability p of "success" (e.g., "heads"), and count the total number of successes.

Example 1.23 [Poisson distribution] Let $\lambda > 0$ and $p_k := \mathrm{e}^{-\lambda}(\lambda^k/k!)$ for $k \in \mathbb{Z}_+$. If $\mathbb{P}[X = k] = p_k$, X is a Poisson random variable with parameter λ.

Let X be a discrete random variable. The *expectation*, *expected value*, or *mean* of X is given by

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\mathbb{P}[X = x],$$

provided the sum is finite. The *variance* of X is $\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Example 1.24 [Indicator variables.] Let A be an event. Let $\mathbf{1}_A$ denote the *indicator random variable* of A, that is, $\mathbf{1}_A : \Omega \to \{0, 1\}$ given by

$$\mathbf{1}_A(\omega) := \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$$

So $\mathbf{1}_A$ is 1 if A happens and 0 if not. Then

$$\begin{aligned}\mathbb{E}[\mathbf{1}_A] &= 1 \cdot \mathbb{P}[\mathbf{1}_A = 1] + 0 \cdot \mathbb{P}[\mathbf{1}_A = 0] \\ &= \mathbb{P}[\mathbf{1}_A = 1] = \mathbb{P}[A].\end{aligned}$$

Expectation has the following basic properties. For X and Y random variables with well-defined expectations and $a, b \in \mathbb{R}$,

(a) $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ (linearity).

(b) If $\mathbb{P}[X \le Y] = 1$, then $\mathbb{E}[X] \le \mathbb{E}[Y]$ (monotonicity).

(c) $|\mathbb{E}[X]| \le \mathbb{E}[|X|]$ (triangle inequality).

(d) If $h : X(\Omega) \to \mathbb{R}$, then $\mathbb{E}[h(X)] = \sum_{x \in X(\Omega)} h(x)\mathbb{P}[X = x]$ ("law of the unconscious statistician").

Let X be a nonnegative random variable. Then, for any $x > 0$, $x\mathbf{1}\{X \ge x\} \le X$ holds with probability 1. Taking expectations and using monotonicity yields $\mathbb{P}[X \ge x] \le x^{-1}\mathbb{E}[X]$. This is usually known as *Markov's inequality*, although it is also sometimes referred to as *Chebyshev's inequality*.[15] Applying Markov's inequality to $e^{\theta X}$, $\theta > 0$, gives

$$\mathbb{P}[X \ge x] = \mathbb{P}[e^{\theta X} \ge e^{\theta x}] \le e^{-\theta x}\mathbb{E}[e^{\theta X}],$$

which is sometimes known as *Chernoff's inequality*.

Let $(\Omega, \mathbb{P})$ be a discrete probability space. A family $(X_i, i \in I)$ of random variables is called *independent* if for any finite subset $J \subseteq I$ and all $x_j \in X_j(\Omega)$,

$$\mathbb{P}\Big(\bigcap_{j \in J} \{X_j = x_j\}\Big) = \prod_{j \in J} \mathbb{P}(X_j = x_j).$$

In particular, random variables X and Y are independent if $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\mathbb{P}[Y = y]$ for all x and y.

Theorem 1.18 *If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

A consequence of Theorem 1.18 is that if X and Y are independent, then $\mathbb{V}\mathrm{ar}[X + Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]$.

Example 1.25

(a) If $Y \sim \mathrm{Be}(p)$ then $\mathbb{E}[Y] = \mathbb{E}[Y^2] = p \cdot 1 + (1 - p) \cdot 0 = p$, so $\mathbb{V}\mathrm{ar}[Y] = p - p^2 = p(1 - p)$.

(b) If $X \sim \mathrm{Bin}(n, p)$ then we can write $X = \sum_{i=1}^n Y_i$ where $Y_i \sim \mathrm{Be}(p)$ are independent. By linearity of expectation, $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[Y_i] = np$. Also, by independence, $\mathbb{V}\mathrm{ar}[X] = \sum_{i=1}^n \mathbb{V}\mathrm{ar}[Y_i] = np(1 - p)$.

1.A.5 Conditional Expectation

On a discrete probability space $(\Omega, \mathbb{P})$, let B be an event with $\mathbb{P}[B] > 0$ and let X be

15) Chebyshev's inequality is the name more commonly associated with Markov's inequality applied to the random variable $(X - \mathbb{E}[X])^2$, to give $\mathbb{P}[|X - \mathbb{E}[X]| \ge x] \le x^{-2}\mathbb{V}\mathrm{ar}[X]$.

a random variable. The *conditional expectation* of X given B is

$$\mathbb{E}[X \mid B] = \sum_{x \in X(\Omega)} x \mathbb{P}[X = x \mid B].$$

So $\mathbb{E}[X \mid B]$ can be thought of as expectation with respect to the conditional probability measure $\mathbb{P}[\,\cdot \mid B]$. An alternative is

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X \mathbf{1}_B]}{\mathbb{P}[B]}, \tag{1.35}$$

where $\mathbf{1}_B$ is the indicator random variable of B. The proof of (1.35) is an exercise in interchanging summations. First,

$$\begin{aligned}\mathbb{E}[X \mid B] &= \sum_{x \in X(\Omega)} x \mathbb{P}[X = x \mid B] \\ &= \sum_{x \in X(\Omega)} x \frac{\mathbb{P}[\{X = x\} \cap B]}{\mathbb{P}[B]}. \end{aligned} \tag{1.36}$$

On the other hand, the random variable $\mathbf{1}_B X$ takes values $x \neq 0$ with

$$\begin{aligned}\mathbb{P}[\mathbf{1}_B X = x] &= \sum_{\omega \in \Omega : \omega \in B \cap \{X = x\}} \mathbb{P}[\{\omega\}] \\ &= \mathbb{P}[\{X = x\} \cap B],\end{aligned}$$

so by comparison we see that the final expression in (1.36) is indeed $\mathbb{E}[\mathbf{1}_B X]/\mathbb{P}[B]$.

Let $(E_i, i \in I)$ be a partition of Ω, so $\sum_{i \in I} \mathbf{1}_{E_i} = 1$. Hence,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\Big[X \sum_{i \in I} \mathbf{1}_{E_i}\Big] = \mathbb{E}\Big[\sum_{i \in I} X \mathbf{1}_{E_i}\Big] \\ &= \sum_{i \in I} \mathbb{E}[X \mathbf{1}_{E_i}],\end{aligned}$$

by linearity. By (1.35), $\mathbb{E}[X \mathbf{1}_{E_i}] = \mathbb{E}[X \mid E_i]\mathbb{P}[E_i]$. Thus we verify the *partition theorem* for expectations:

$$\mathbb{E}[X] = \sum_{i \in I} \mathbb{E}[X \mid E_i]\mathbb{P}[E_i].$$

Given two discrete random variables X and Y, the *conditional expectation* of X given Y, denoted $\mathbb{E}[X \mid Y]$, is the *random variable* $\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = Y(\omega)]$, which takes values $\mathbb{E}[X \mid Y = y]$ with probabilities $\mathbb{P}[Y = y]$.

References

1. Durrett, R. (2010) *Probability: Theory and Examples*, 4th edn, Cambridge University Press, Cambridge.
2. Kallenberg, O. (2002) *Foundations of Modern Probability*, 2nd edn, Springer-Verlag, New York.
3. Fischer, H. (2011) *A History of the Central Limit Theorem*, Springer-Verlag, New York.
4. Ehrenfest, P. and Ehrenfest, T. (1907) Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. *Phys. Z.*, **8**, 311–314.
5. Karlin, S. and Taylor, H.M. (1975) *A First Course in Stochastic Processes*, 2nd edn, Academic Press.
6. Norris, J.R. (1997) *Markov Chains*, Cambridge University Press, Cambridge.
7. Chung, K.L. (1967) *Markov Chains with Stationary Transition Probabilities*, 2nd edn, Springer-Verlag, Berlin.
8. Lamperti, J. (1960) Criteria for the recurrence or transience of stochastic process. I. *J. Math. Anal. Appl.*, **1**, 314–330.
9. Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd edn, John Wiley & Sons, Inc., New York.
10. Bak, P. and Sneppen, K. (1993) Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.*, **71**, 4083–4086.
11. Gillett, A.J. (2007) Phase Transitions in Bak–Sneppen Avalanches and in a Continuum Percolation Model. PhD dissertation, Vrije Universiteit, Amsterdam.
12. Grinfeld, M., Knight, P.A., and Wade, A.R. (2012) Rank-driven Markov processes. *J. Stat. Phys.*, **146**, 378–407.
13. Rayleigh, L. (1880) On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *Philos. Mag.*, **10**, 73–78.
14. Bachelier, L. (1900) Théorie de la spéculation. *Ann. Sci. École Norm. Sup.*, **17**, 21–86.

15. Pearson, K. and Blakeman, J. (1906) *A Mathematical Theory of Random Migration*, *Drapers' Company Research Memoirs Biometric Series*, Dulau and co., London.
16. Einstein, A. (1956) *Investigations on the Theory of the Brownian Movement*, Dover Publications Inc., New York.
17. Pólya, G. (1921) Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. *Math. Ann.*, **84**, 149–160.
18. Pearson, K. (1905) The problem of the random walk. *Nature*, **72**, 342.
19. Hughes, B.D. (1995) *Random Walks and Random Environments*, vol. 1, Oxford University Press, New York.
20. Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, Vol. I, 3rd edn, John Wiley & Sons, Inc., New York.
21. Doherty, M. (1975) *An Amusing Proof in Fluctuation Theory*, *Combinatorial Mathematics III*, vol. 452, Springer, 101–104.
22. Ganesh, A., O'Connell, N., and Wischik, D. (2004) *Big Queues*, Springer, Berlin.
23. Foss, S., Korshunov, D., and Zachary, S. (2011) *An Introduction to Heavy-Tailed and Subexponential Distributions*, Springer-Verlag.
24. Dembo, A. and Zeitouni, O. (1998) *Large Deviations Techniques and Applications*, Springer, New York.
25. Touchette, H. (2009) The large deviation approach to statistical mechanics. *Phys. Rep.*, **478**, 1–69.
26. Karlin, S. and Taylor, H.M. (1981) *A Second Course in Stochastic Processes*, Academic Press.
27. Lax, M., Cai, W., and Xu, M. (2006) *Random Processes in Physics and Finance*, Oxford University Press.
28. Blackman, J.A. and Mulheran, P.A. (1996) Scaling behaviour in submonolayer film growth: a one-dimensional model. *Phys. Rev. B*, **54**, 11 681.
29. O'Neill, K.P., Grinfeld, M., Lamb, W., and Mulheran, P.A. (2012) Gap-size and capture-zone distributions in one-dimensional point-island nucleation and growth simulations: Asymptotics and models. *Phys. Rev. E*, **85**, 21 601.
30. Ising, E. (1925) Beitrag zur Theorie des Ferromagnetismus. *Z. Phys.*, **31**, 253–258.
31. Grimmett, G. (2010) *Probability on Graphs*, Cambridge University Press.
32. Lawler, G.F. (2005) *Conformally Invariant Processes in the Plane*, American Mathematical Society.
33. Grimmett, G. (1999) *Percolation*, 2nd edn, Springer-Verlag, Berlin.
34. Doyle, P.G. and Snell, J.L. (1984) *Random Walks and Electric Networks*, Mathematical Association of America, Washington, DC.
35. Kac, M. (1959) *Probability and Related Topics in Physical Sciences*, vol. 1957, Interscience Publishers, London and New York.
36. Shlesinger, M.F. and West, B.J. (1984) *Random Walks and Their Applications on the Physical and Biological Sciences*, American Institute of Physics, New York.
37. Weiss, G.H. and Rubin, R.J. (1983) Random walks: theory and selected applications. *Adv. Chem. Phys.*, **52**, 363–505.
38. Streater, R.F. (2007) *Lost Causes in and Beyond Physics*, Springer-Verlag.
39. Billingsley, P. (1995) *Probability and Measure*, 3rd edn, John Wiley & Sons, Inc. New York.
40. Chung, K.L. (2001) *A Course in Probability Theory*, 3rd edn, Academic Press Inc., San Diego, CA.