3 Random Variables

The material in future chapters will require familiarity with basic concepts from probability theory, random variables, and random processes. For the benefit of the reader, we review in this and the next chapters several concepts of general interest. The discussion is not meant to be comprehensive or exhaustive. Only concepts that are necessary for our treatment of inference and learning are reviewed. It is assumed that readers have had some prior exposure to random variables and probability theory.

3.1 PROBABILITY DENSITY FUNCTIONS

In loose terms, the designation "random variable" refers to a variable whose value cannot be predicted with certainty prior to observing it. This is because the variable may assume any of a collection of values in an experiment, and some of the values can be more likely to occur than other values. In other words, there is an element of *chance* associated with each possibility.

In our treatment, we will often (but not exclusively) be interested in random variables whose observations assume *numerical* values. Obviously, in many situations of interest, the random variables need not be numerical but are categorical in nature. One example is when a ball is drawn from an urn and is either blue-colored with probability $1/4$ or red-colored with probability $3/4$. In this case, the qualifications $\{\text{red}, \text{blue}\}$ refer to the two possible outcomes, which are not numerical. Nevertheless, it is common practice in scenarios like this to associate numerical values with each category, such as assigning the numerical value $+1$ to the color red and the numerical value -1 to the color blue. In this way, drawing a red ball amounts to observing the value $+1$ with probability $3/4$ and drawing a blue ball amounts to observing the value -1 with probability $1/4$.

We will use **boldface** symbols to refer to random variables and symbols in *normal* font to refer to their realizations or observations. For example, we let $\boldsymbol{x}$ denote the random variable that corresponds to the outcome of throwing a dice. Each time the experiment is repeated, one of six possible outcomes can be observed, namely, $x \in \{1, 2, 3, 4, 5, 6\}$ — see Fig. 3.1. We cannot tell beforehand which value will occur (assuming a fair dice). We say that the random variable, $\boldsymbol{x}$, represents the outcome of the experiment and each observation x is a realization

for $\boldsymbol{x}$. In this example, the realization x can assume one of six possible integer values, which constitute the *sample space* for $\boldsymbol{x}$ and is denoted by the letter $\Omega = \{1, 2, 3, 4, 5, 6\}$. Our choice of notation $\{\boldsymbol{x}, x\}$ is meant to distinguish between a random variable and its realizations or observations.

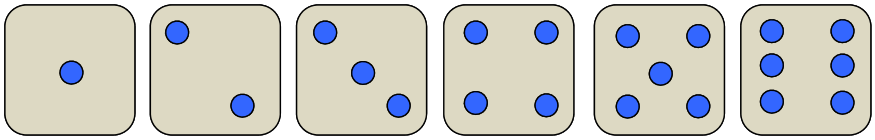

Figure 3.1 The sample space for a dice consists of the outcomes $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Discrete variables

A numerical random variable can be *discrete* or *continuous* depending on the range of values it assumes. The realizations of a discrete random variable can only assume values from a countable set of distinct possibilities. For example, the outcome of the throw of a dice is a discrete random variable since the realizations can only assume one of six possible values from the set $\Omega = \{1, 2, 3, 4, 5, 6\}$. In contrast, the realizations of a continuous random variable can assume an infinite number of possible values, e.g., values within an interval on the real line.

When a random variable is discrete, we associate with each element of the sample space, Ω, a nonnegative number in the range $[0, 1]$. This number represents the probability of occurrence of that particular element of Ω. For example, assuming a fair dice throw, the probability that the realization $x = 4$ is observed is equal to $1/6$. This is because all six possible outcomes are equally likely with the same probability of occurrence, which we denote by

$$p_m = 1/6, \quad m = 1, 2, 3, 4, 5, 6 \tag{3.1}$$

with one value p_m for each possible outcome m. Obviously, the sum of all six probabilities must add up to one. We refer to the $\{p_m\}$ as representing the *probability distribution* or the *probability mass function* (pmf) that is associated with the dice experiment. More generally, for a discrete random variable, $\boldsymbol{x}$, with M possible realizations, $\{x_m\}$, we associate with each outcome a probability value p_m for all $1 \leq m \leq M$. These probabilities need not be identical because some outcomes may be more likely to occur than others, but they must satisfy the following two conditions:

$$\boxed{0 \leq p_m \leq 1 \quad \text{and} \quad \sum_{m=1}^{M} p_m = 1} \tag{3.2}$$

with the number p_m corresponding to the probability of the $m-$th event occur-

ring, written as

$$p_m \triangleq \mathbb{P}(\boldsymbol{x} = x_m) \tag{3.3}$$

When convenient, we will also use the alternative function notation $f_{\boldsymbol{x}}(x_m)$ to refer to the probability of event x_m, namely,

$$f_{\boldsymbol{x}}(x_m) \triangleq \mathbb{P}(\boldsymbol{x} = x_m) = p_m \tag{3.4}$$

where $f_{\boldsymbol{x}}(x)$ refers to a function that assumes the value p_m at each location x_m, and the value zero at all other locations. More formally, $f_{\boldsymbol{x}}(x)$ can be expressed in terms of the Dirac delta function, $\delta(x)$, as follows:

$$f_{\boldsymbol{x}}(x) = \sum_{m=1}^{M} p_m \, \delta(x - x_m) \tag{3.5}$$

where the delta function is defined by the sifting property:

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_m)dx \;=\; g(x_m) \tag{3.6}$$

for any function $g(x)$ that is well-defined at $x = x_m$. Representation (3.5) highlights the fact that the probability distribution of a discrete random variable, $\boldsymbol{x}$, is concentrated at a finite number of locations defined by the coordinates $\{x_m\}$.

Continuous variables

The function notation $f_{\boldsymbol{x}}(x)$ for the pmf of a discrete random variable is useful because, as explained next, it will allow us to adopt a common notation for both discrete and continuous random variables.

When the random variable $\boldsymbol{x}$ is continuous, the probability that $\boldsymbol{x}$ assumes any particular value x from its sample space is equal to zero. This is because there are now infinitely many possible realization values. For this reason, for continuous random variables, we are more interested in the probability of events involving a *range* of values rather than a specific value. To evaluate the probability of such events, we associate with the random variable $\boldsymbol{x}$ a *probability density function* (pdf), which we will denote by the same notation $f_{\boldsymbol{x}}(x)$. The pdf is a function of x and it is required to satisfy the following two conditions:

$$\boxed{f_{\boldsymbol{x}}(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x)dx = 1} \tag{3.7}$$

The pdf of $\boldsymbol{x}$ allows us to evaluate probabilities of events of the form

$$\mathbb{P}(a \leq \boldsymbol{x} \leq b) \tag{3.8}$$

which refer to the probability that $\boldsymbol{x}$ assumes values within the interval $[a, b]$. This probability is obtained through the integral calculation:

$$\mathbb{P}(a \leq \boldsymbol{x} \leq b) \;=\; \int_{a}^{b} f_{\boldsymbol{x}}(x)dx \tag{3.9}$$

We will use the terminology of "probability *mass functions*" for discrete random variables, and "probability *density functions*" for continuous random variables. Moreover, we will often use the same pdf notation, $f_{\boldsymbol{x}}(x)$, to refer to probability distributions in both cases.

Example 3.1 (**Uniform random variable**) A continuous random variable $\boldsymbol{x}$ is said to be uniformly distributed within the interval $[a, b]$ if its pdf is constant over this interval and zero elsewhere, namely,

$$f_{\boldsymbol{x}}(x) = \begin{cases} c, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

for some constant $c > 0$ and where $b > a$. The value of c can be determined from the normalization requirement

$$\int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x)dx = 1 \tag{3.11}$$

so that we must have

$$\int_{a}^{b} c\, dx = 1 \implies c = \frac{1}{b-a} \tag{3.12}$$

We conclude that the pdf of a uniform random variable is given by

$$f_{\boldsymbol{x}}(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

3.2 MEAN AND VARIANCE

Consider a continuous real-valued random variable $\boldsymbol{x}$ and let $x \in \mathcal{X}$ denote the domain of its realizations (i.e., the range of values that can be assumed by x). For example, for the uniform variable described in Example 3.1, we have $\mathcal{X} = [a\ b]$.

Definitions

The mean $\boldsymbol{x}$ is denoted by $\bar{x}$ or $\mathbb{E}\,\boldsymbol{x}$ and is defined as the calculation:

$$\mathbb{E}\,\boldsymbol{x} \triangleq \bar{x} = \int_{x\in\mathcal{X}} x f_{\boldsymbol{x}}(x)dx \tag{3.14}$$

The mean of $\boldsymbol{x}$ is also called the expected value or the first-moment of $\boldsymbol{x}$, and its computation can be interpreted as determining the center of mass of the pdf. This interpretation is illustrated by the following example.

Likewise, the variance of a real-valued random variable $\boldsymbol{x}$ is denoted by σ_x^2 and defined by the following equivalent expressions:

$$\sigma_x^2 \triangleq \mathbb{E}\,(\boldsymbol{x} - \bar{x})^2 = \int_{x\in\mathcal{X}} (x-\bar{x})^2 f_{\boldsymbol{x}}(x)dx \tag{3.15a}$$

$$= \mathbb{E}\,\boldsymbol{x}^2 - \bar{x}^2 = \left(\int_{x\in\mathcal{X}} x^2 f_{\boldsymbol{x}}(x)dx\right) - \bar{x}^2 \tag{3.15b}$$

Obviously, the variance is a nonnegative number,

$$\boxed{\sigma_x^2 \geq 0} \tag{3.16}$$

and its square-root, which we denote by σ_x, is referred to as the *standard deviation* of $\boldsymbol{x}$. When $\boldsymbol{x}$ has zero mean, it is seen from (3.15a) that its variance expression reduces to the second-order moment of $\boldsymbol{x}$, i.e.,

$$\sigma_x^2 = \mathbb{E}\,\boldsymbol{x}^2 \;=\; \int_{-\infty}^{\infty} x^2 f_{\boldsymbol{x}}(x)dx, \qquad \text{when } \mathbb{E}\,\boldsymbol{x} = 0 \tag{3.17}$$

Example 3.2 (**Mean and variance of a uniform random variable**) Let us reconsider the uniform pdf from Example 3.1. The mean of $\boldsymbol{x}$ is given by

$$\bar{x} = \int_a^b x\,\frac{1}{b-a}dx \;=\; \frac{1}{b-a}\frac{x^2}{2}\bigg|_a^b \;=\; \frac{1}{2}(a+b) \tag{3.18}$$

which is the midpoint of the interval $[a, b]$. The variance of $\boldsymbol{x}$ is given by

$$\begin{aligned}\sigma_x^2 &= \int_a^b x^2\,\frac{1}{b-a}dx \;-\; \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{b-a}\frac{x^3}{3}\bigg|_a^b \;-\; \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{12}(b-a)^2\end{aligned} \tag{3.19}$$

Example 3.3 (**Center of mass**) Consider a rod of length ℓ and unit-mass lying horizontally along the $x-$axis. The left-end of the rod is the origin of the horizontal axis, and the distribution of mass density across the rod is described by the function $f_{\boldsymbol{x}}(x)$ (measured in mass/unit length). Specifically, the mass content between locations x_1 and x_2 is given by the integral of $f_{\boldsymbol{x}}(x)$ over the interval $[x_1, x_2]$. The unit-mass assumption means that

$$\int_0^{\ell} f_{\boldsymbol{x}}(x)dx = 1 \tag{3.20}$$

If left unattended, the rod will swing around its left-end. We would like to determine the $x-$coordinate of the center of mass of the rod where it can be stabilized. We denote this location by $\bar{x}$. The mass of the rod to the left of $\bar{x}$ exerts a torque that would make the rod rotate in an anti-clockwise direction, while the mass of the rod to the right of $\bar{x}$ exerts a torque that would make the rod rotate in a clockwise direction — see Fig. 3.2. An equilibrium is reached when these two torques are balanced against each other. Recall that torque is force multiplied by distance and the forces present are the cumulative weights of the respective parts of the rod to the left and right of $\bar{x}$. Therefore, the equilibrium condition amounts to:

$$\int_0^{\bar{x}} (\bar{x} - x) g f_{\boldsymbol{x}}(x)dx \;=\; \int_{\bar{x}}^{\ell} (x - \bar{x}) g f_{\boldsymbol{x}}(x)dx \tag{3.21}$$

where g is the gravitational acceleration constant, approximately equal to $g = 9.8\,\mathrm{m/s^2}$. Solving for $\bar{x}$ we find that

$$\bar{x} = \int_0^{\ell} x f_{\boldsymbol{x}}(x)dx \tag{3.22}$$

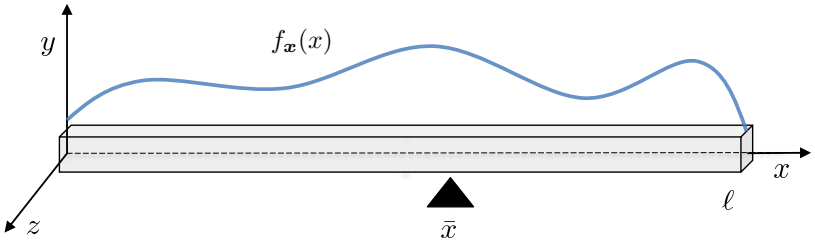

Figure 3.2 A rod of unit-mass and length ℓ is balanced horizontally at location $\bar{x}$.

Although the previous definitions for mean and variance assume a continuous random variable, they are also applicable to discrete random variables if we resort to the representation (3.5). Indeed, in this case, and assuming M possible outcomes $\{x_m\}$, each with a probability of occurrence p_m, the mean and variance relations (3.14) and (3.15b) simplify to the following expressions where integrals are replaced by sums:

$$\bar{x} = \sum_{m=1}^{M} p_m x_m \tag{3.23}$$

$$\sigma_x^2 = \left(\sum_{m=1}^{M} p_m x_m^2 \right) - \bar{x}^2 \tag{3.24}$$

Measure of uncertainty

The variance of a random variable admits a useful interpretation as a measure of uncertainty. Intuitively, the variance σ_x^2 defines an interval on the real axis around the mean $\bar{x}$ where the values of the random variable $\boldsymbol{x}$ are most likely to occur:

(a) A small value of σ_x^2 indicates that $\boldsymbol{x}$ is more likely to assume values close to its mean, $\bar{x}$. In this case, we would have a reasonably good idea about what range of values are likely to be observed for $\boldsymbol{x}$ in experiments.

(b) A large value of σ_x^2 indicates that $\boldsymbol{x}$ can assume values over a wider interval around its mean. In this case, we are less certain about what values to expect for $\boldsymbol{x}$ in experiments.

For this reason, it is customary to regard the variance of a random variable as a measure of the *uncertainty* about the value it will assume in a given experiment. A small variance indicates that we are more certain about what values to expect for $\boldsymbol{x}$ (namely, values that are close to its mean), while a large variance indicates that we are less certain about what values to expect. These two situations are

illustrated in Figs. 3.3 and 3.4 for two different probability density functions.

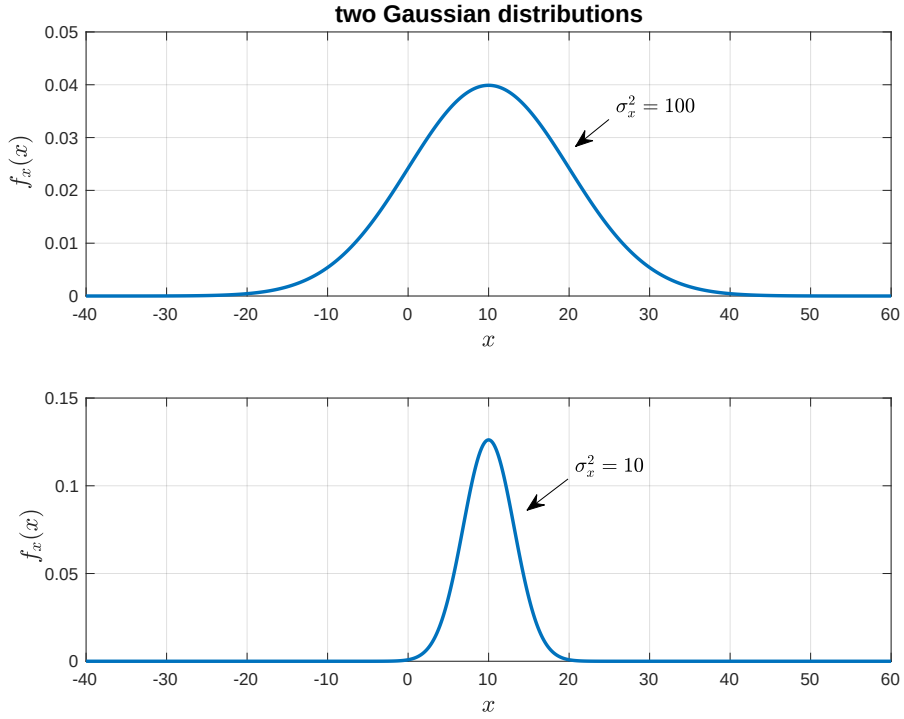

Figure 3.3 Probability density functions $f_{\boldsymbol{x}}(x)$ of a Gaussian random variable $\boldsymbol{x}$ with mean $\bar{x} = 10$, variance $\sigma_x^2 = 100$ in the top plot, and variance $\sigma_x^2 = 10$ in the bottom plot.

Figure 3.3 plots the probability density function of a Gaussian-distributed random variable $\boldsymbol{x}$ for two different variances. In both cases, the mean of the random variable is fixed at $\bar{x} = 10$ while the variance is $\sigma_x^2 = 100$ in one case and $\sigma_x^2 = 10$ in the other. We explain Chapter 4 that the pdf of a Gaussian random variable is defined in terms of $(\bar{x}, \sigma_x^2)$ by the following expression — see (4.4):

$$\boxed{f_{\boldsymbol{x}}(x) \;=\; \frac{1}{\sqrt{2\pi}\;\sigma_x}\; e^{-(x-\bar{x})^2/2\sigma_x^2}, \quad x \in (-\infty, \infty)} \qquad (\textbf{Gaussian}) \qquad (3.25)$$

From Fig. 3.3 we observe that the smaller the variance of $\boldsymbol{x}$ is, the more concentrated its pdf is around its mean. Figure 3.4 provides similar plots for a second random variable $\boldsymbol{x}$ with a Rayleigh distribution, namely, with a pdf given by

$$\boxed{f_{\boldsymbol{x}}(x) \;=\; \frac{x}{\alpha^2}\; e^{-x^2/2\alpha^2}, \quad x \geq 0, \;\; \alpha > 0} \qquad (\textbf{Rayleigh}) \qquad (3.26)$$

where α is a positive parameter. The value of α determines the mean and variance

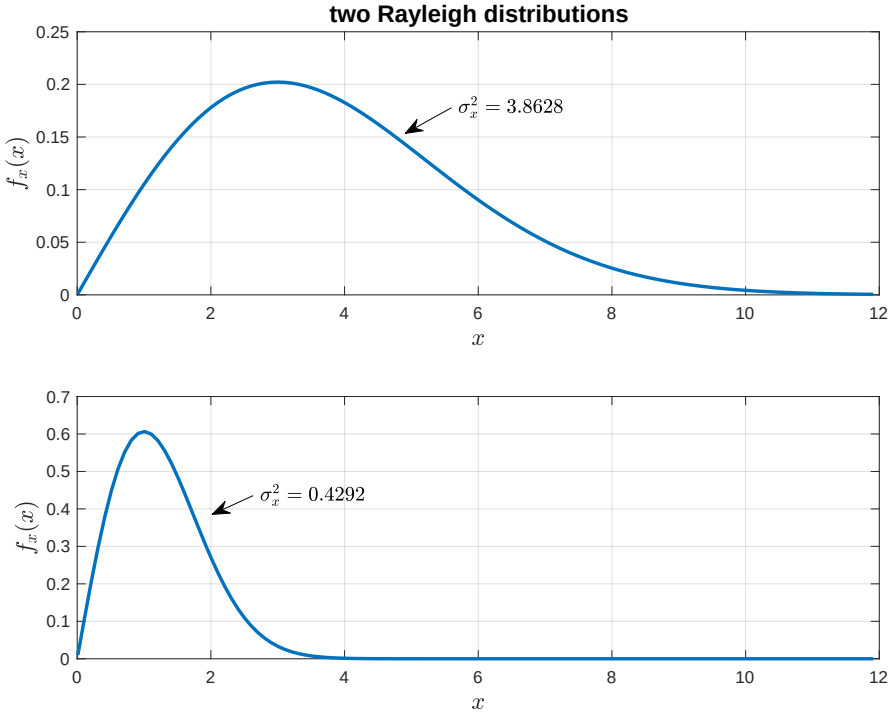

Figure 3.4 Probability density functions $f_{\boldsymbol{x}}(x)$ of a Rayleigh random variable $\boldsymbol{x}$ with mean $\bar{x} = 3.7599$ and variance $\sigma_x^2 = 3.8628$ in the top plot, and mean $\bar{x} = 1.2533$ and variance $\sigma_x^2 = 0.4292$ in the bottom plot.

of $\boldsymbol{x}$ according to the following expressions (see Prob. 3.15):

$$\bar{x} = \alpha \sqrt{\frac{\pi}{2}}, \qquad \sigma_x^2 = \left(2 - \frac{\pi}{2}\right) \alpha^2 \tag{3.27}$$

so that, in contrast to the Gaussian case, the mean and variance of a Rayleigh-distributed random variable cannot be chosen independently of each other since their values are linked through α. In Fig. 3.4, the top plot corresponds to $\bar{x} = 3.7599$ and $\sigma_x^2 = 3.8628$, while the bottom plot corresponds to $\bar{x} = 1.2533$ and $\sigma_x^2 = 0.4292$.

Chebyshev inequality

The above remarks on the variance of a random variable can be qualified more formally by invoking a well-known result from probability theory known as *Chebyshev inequality* — see Probs. 3.17 and 3.18. The result states that for a random variable $\boldsymbol{x}$ with mean $\bar{x}$ and finite variance σ_x^2, and for any given scalar $\delta > 0$, it holds that

$$\boxed{\mathbb{P}(|\boldsymbol{x} - \bar{x}| \geq \delta) \;\; \leq \;\; \sigma_x^2/\delta^2} \tag{3.28}$$

This inequality is meaningful only for values of δ satisfying $\delta \geq \sigma_x$; otherwise, the right-hand side becomes larger than one and the inequality becomes trivial Result (3.28) states that the probability that $\boldsymbol{x}$ assumes values outside the interval $(\bar{x}-\delta, \bar{x}+\delta)$ does not exceed σ_x^2/δ^2, with the bound being proportional to the variance of $\boldsymbol{x}$. Hence, for a fixed δ, the smaller the variance of $\boldsymbol{x}$ is, the smaller the probability that $\boldsymbol{x}$ will assume values outside the interval $(\bar{x}-\delta, \bar{x}+\delta)$. If δ is selected as a multiple of the standard deviation of $\boldsymbol{x}$, say, as $\delta = q\sigma_x$, for some $q \geq 1$, then we conclude from Chebyshev inequality that

$$\mathbb{P}(|\boldsymbol{x}-\bar{x}| \geq q\sigma_x) \;\leq\; 1/q^2, \quad q \geq 1 \tag{3.29}$$

Let us choose, for example, $\delta = 5\sigma_x$. Then, expression (3.29) gives

$$\mathbb{P}(|\boldsymbol{x}-\bar{x}| \geq 5\sigma_x) \;\leq\; 1/25 = 4\% \tag{3.30}$$

In other words, there is at most 4% chance that $\boldsymbol{x}$ will assume values outside the interval $(\bar{x}-5\sigma_x, \bar{x}+5\sigma_x)$. Actually, the bound that is provided by Chebyshev inequality is generally loose, as the following example illustrates.

Example 3.4 (**Gaussian case**) Consider a zero-mean Gaussian random variable $\boldsymbol{x}$ with variance σ_x^2 and choose $\delta = 2\sigma_x$. Then, from Chebyshev inequality (3.29) we obtain

$$\mathbb{P}(|\boldsymbol{x}| \geq 2\sigma_x) \;\leq\; 1/4 = 25\% \tag{3.31}$$

whereas direct evaluation of the probability using the Gaussian pdf (3.25) gives

$$\begin{aligned}
\mathbb{P}(|\boldsymbol{x}| \geq 2\sigma_x) &= \mathbb{P}(\boldsymbol{x} \geq 2\sigma_x) \;+\; \mathbb{P}(\boldsymbol{x} \leq -2\sigma_x) \\
&= \frac{1}{\sqrt{2\pi}\,\sigma_x}\left(\int_{2\sigma_x}^{\infty} e^{-x^2/2\sigma_x^2}\,dx \;+\; \int_{-\infty}^{-2\sigma_x} e^{-x^2/2\sigma_x^2}\,dx\right) \\
&= 1 - 2\left(\frac{1}{\sqrt{2\pi}\,\sigma_x}\int_{0}^{2\sigma_x} e^{-x^2/2\sigma_x^2}\,dx\right)
\end{aligned} \tag{3.32}$$

which can be evaluated numerically to yield:

$$\mathbb{P}(|\boldsymbol{x}| \geq 2\sigma_x) \;\approx\; 4.56\% \tag{3.33}$$

Example 3.5 (**Zero-variance random variables**) One useful consequence of Chebyshev inequality (3.28) is that it allows us to interpret a zero-variance random variable as one that is equal to its mean in probability — see also Prob. 3.42. This is because when $\sigma_x^2 = 0$, we obtain from (3.28) that for *any* small $\delta > 0$:

$$\mathbb{P}(|\boldsymbol{x}-\bar{x}| \geq \delta) \;\leq\; 0 \tag{3.34}$$

But since the probability of any event is necessarily a nonnegative number, we conclude that

$$\mathbb{P}(|\boldsymbol{x}-\bar{x}| \geq \delta) = 0, \;\; \text{for any } \delta > 0 \tag{3.35}$$

We say in this case that the equality $\boldsymbol{x} = \bar{x}$ holds in probability:

$$\boxed{\sigma_x^2 = 0 \;\Longrightarrow\; \boldsymbol{x} = \bar{x} \quad \text{in probability}} \tag{3.36}$$

For the benefit of the reader, we explain in Appendix 3.A various notions of convergence for random variables, including convergence in probability, convergence in distribution,

almost-sure convergence, and mean-square convergence. For the current example, the convergence in probability result (3.36) is equivalent to statement (3.35).

3.3 DEPENDENT RANDOM VARIABLES

In inference problems, it is generally the case that information about one unobservable random variable is inferred from observations of another random variable. The observations of the second random variable will convey more or less information about the desired variable depending on how closely related (i.e., dependent) the two random variables are. Let us illustrate this concept using two examples.

Example 3.6 (**Independent random variables**) Assume a dice is rolled twice. Let $\boldsymbol{x}$ denote the random variable that represents the outcome of the first roll and let $\boldsymbol{z}$ denote the random variable that represents the outcome of the second roll. These two random variables are independent of each other since the outcome of one experiment does not influence the outcome of the other. For example, if it is observed that $z = 5$, then this value does not tell us anything about what value $\boldsymbol{x}$ assumed in the first roll. Likewise, if $x = 4$, then this value does not tell us anything about what value $\boldsymbol{z}$ will assume in the second roll. That is, observations of one variable do not provide any information about the other variable.

Example 3.7 (**Two throws of a dice**) Assume again that the dice is rolled twice. Let $\boldsymbol{x}$ denote the random variable that represents the outcome of the first roll. Let $\boldsymbol{y}$ denote the random variable that represents the *sum* of the two rolls. Assume we only observe the outcome of $\boldsymbol{y}$ and are unaware of the outcome of $\boldsymbol{x}$. Obviously, the observation of $\boldsymbol{y}$ conveys some information about $\boldsymbol{x}$. For example, if the observation of $\boldsymbol{y}$ is $y = 10$, then x could not be $1, 2$, or 3 because in these cases the result of the second roll can never result in a sum that is equal to 10. We therefore say that the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent (the value assumed by one variable in a given experiment conveys some information about the potential value assumed by the other variable). When random variables are dependent in this way, it becomes possible to use observations of one variable to infer the value of the other random variable. Obviously, the result of the estimation is generally imperfect and it is rarely the case that we can infer precisely the value of the unobserved variable. In most situations, we will be satisfied with close enough guesses, where the measure of "closeness" will be formalized in some well-defined manner, e.g., by using the mean-square-error criterion or other criteria.

3.3.1 Bayes Rule

The dependency between two real-valued random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ is captured by their *joint* probability density function, which is a two-dimensional function denoted by $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$. The joint pdf allows us to evaluate probabilities of events of the form:

$$\mathbb{P}(a \leq \boldsymbol{x} \leq b, c \leq \boldsymbol{y} \leq d) = \int_c^d \int_a^b f_{\boldsymbol{x},\boldsymbol{y}}(x, y) dx dy \tag{3.37}$$

namely, the probability that $\boldsymbol{x}$ and $\boldsymbol{y}$ assume values inside the intervals $[a, b]$ and $[c, d]$, respectively. We also introduce the *conditional* pdf of $\boldsymbol{x}$ given $\boldsymbol{y}$, which is denoted by $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$; this function allows us to evaluate probabilities of events of the form:

$$\mathbb{P}(a \leq \boldsymbol{x} \leq b \mid \boldsymbol{y} = y) = \int_a^b f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \tag{3.38}$$

namely, the probability that $\boldsymbol{x}$ assumes values inside the interval $[a, b]$ given that $\boldsymbol{y}$ assumes the value y. It is a well-known result in probability theory that the joint and conditional pdfs of two random variables are related via Bayes rule, which states that:

$$\begin{array}{l}(\boldsymbol{x} \textbf{ and } \boldsymbol{y} \textbf{ are continuous}) \\ f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = f_{\boldsymbol{y}}(y)\ f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = f_{\boldsymbol{x}}(x)\ f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)\end{array} \tag{3.39}$$

This relation expresses the joint pdf as the product of the individual and conditional probability density functions of $\boldsymbol{x}$ and $\boldsymbol{y}$; it also implies that

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \ = \ \frac{f_{\boldsymbol{x}}(x)\ f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)}{f_{\boldsymbol{y}}(y)} \tag{3.40}$$

This relation will arise frequently in our study of inference problems, so much so that the different terms in this expression have their own terminology:

$$f_{\boldsymbol{y}}(y) \text{ is called the } \textbf{evidence} \text{ of } \boldsymbol{y} \tag{3.41a}$$

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \text{ is called the } \textbf{likelihood} \text{ of } \boldsymbol{y} \tag{3.41b}$$

$$f_{\boldsymbol{x}}(x) \text{ is called the } \textbf{prior} \text{ of } \boldsymbol{x} \tag{3.41c}$$

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \text{ is called the } \textbf{posterior} \text{ of } \boldsymbol{x} \tag{3.41d}$$

In inference problems, we will deal frequently with the problem of observing realizations for some random variable and using them to infer the values for some other hidden or unobservable variable. Usually, the notation $\boldsymbol{y}$ plays the role of the observation and $f_{\boldsymbol{y}}(y)$ refers to its pdf, which is also called its *evidence*. The evidence provides information about the distribution of the observations. Likewise, the notation $\boldsymbol{x}$ plays the role of the hidden variable we are interested in estimating or learning about. Its distribution $f_{\boldsymbol{x}}(x)$ is called the *prior*: it represents the distribution of $\boldsymbol{x}$ prior to observing $\boldsymbol{y}$. In the same token, the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ is called the *posterior* because it represents the distribution of $\boldsymbol{x}$ after observing $\boldsymbol{y}$. The second conditional pdf $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$ is called the *likelihood* of $\boldsymbol{y}$ because it shows how likely the values of $\boldsymbol{y}$ are if $\boldsymbol{x}$ were known. The likelihood can also be interpreted as representing a model for the generation of the observation $\boldsymbol{y}$ from knowledge of $\boldsymbol{x}$.

Form (3.39) for Bayes rule assumes that both variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are continuous. There are variations of this rule when one or both of the random variables happen

to be discrete, namely,

($\boldsymbol{x}$ **continuous,** $\boldsymbol{y}$ **discrete**)

$$\begin{aligned} f_{\boldsymbol{x},\boldsymbol{y}}(x,y) &= \mathbb{P}(\boldsymbol{y}=y)\, f_{\boldsymbol{x}|\boldsymbol{y}}(x|\boldsymbol{y}=y) \\ &= f_{\boldsymbol{x}}(x)\, \mathbb{P}(\boldsymbol{y}=y|\boldsymbol{x}=x) \end{aligned} \tag{3.42a}$$

($\boldsymbol{x}$ **discrete,** $\boldsymbol{y}$ **continuous**)

$$\begin{aligned} f_{\boldsymbol{x},\boldsymbol{y}}(x,y) &= \mathbb{P}(\boldsymbol{x}=x)\, f_{\boldsymbol{y}|\boldsymbol{x}}(y|\boldsymbol{x}=x) \\ &= f_{\boldsymbol{y}}(y)\, \mathbb{P}(\boldsymbol{x}=x|\boldsymbol{y}=y) \end{aligned} \tag{3.42b}$$

($\boldsymbol{x}$ **and** $\boldsymbol{y}$ **discrete**)

$$\begin{aligned} \mathbb{P}(\boldsymbol{x}=x,\boldsymbol{y}=y) &= \mathbb{P}(\boldsymbol{y}=y)\, \mathbb{P}(\boldsymbol{x}=x|\boldsymbol{y}=y) \\ &= \mathbb{P}(\boldsymbol{x}=x)\, \mathbb{P}(\boldsymbol{y}=y|\boldsymbol{x}=x) \end{aligned} \tag{3.42c}$$

Table 3.1 summarizes these variations. We will continue our discussion by considering form (3.39) for continuous random variables, but note that the conclusions can be easily extended to combinations of discrete and continuous variables.

Table 3.1 Different forms of Bayes rule depending on the discrete or continuous nature of the random variables.

$\boldsymbol{x}$	$\boldsymbol{y}$	**Bayes rule**
continuous	continuous	$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{y}}(y)\, f_{\boldsymbol{x}\|\boldsymbol{y}}(x\|y)$ $f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{x}}(x)\, f_{\boldsymbol{y}\|\boldsymbol{x}}(y\|x)$
discrete	discrete	$\mathbb{P}(\boldsymbol{x}=x,\boldsymbol{y}=y) = \mathbb{P}(\boldsymbol{y}=y)\, \mathbb{P}(\boldsymbol{x}=x\|\boldsymbol{y}=y)$ $\mathbb{P}(\boldsymbol{x}=x,\boldsymbol{y}=y) = \mathbb{P}(\boldsymbol{x}=x)\, \mathbb{P}(\boldsymbol{y}=y\|\boldsymbol{x}=x)$
discrete	continuous	$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \mathbb{P}(\boldsymbol{x}=x)\, f_{\boldsymbol{y}\|\boldsymbol{x}}(y\|\boldsymbol{x}=x)$ $f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{y}}(y)\, \mathbb{P}(\boldsymbol{x}=x\|\boldsymbol{y}=y)$
continuous	discrete	$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \mathbb{P}(\boldsymbol{y}=y)\, f_{\boldsymbol{x}\|\boldsymbol{y}}(x\|\boldsymbol{y}=y)$ $f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{x}}(x)\, \mathbb{P}(\boldsymbol{y}=y\|\boldsymbol{x}=x)$

Example 3.8 (**Observation under additive noise**) Assume $\boldsymbol{x}$ is a discrete random variable that assumes the values ± 1 with probability p for $+1$ and $1-p$ for -1. Assume further that $\boldsymbol{v}$ is a zero-mean Gaussian random variable with variance σ_v^2. In a given experiment, the user observes a scaled version of $\boldsymbol{x}$ in the presence of the noise variable $\boldsymbol{v}$. Specifically, the user observes the random variable

$$\boldsymbol{y} = \frac{1}{2}\boldsymbol{x} + \boldsymbol{v} \tag{3.43}$$

Clearly, the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent since realizations of $\boldsymbol{x}$ alter the pdf of $\boldsymbol{y}$. For example, if $x = +1$, then the random variable $\boldsymbol{y}$ will be Gaussian-distributed with mean $+\frac{1}{2}$ and variance σ_v^2, written as

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|x=+1) \;=\; \mathcal{N}_{\boldsymbol{y}}\left(1/2, \sigma_v^2\right) \tag{3.44}$$

where the notation $\mathcal{N}_{\boldsymbol{a}}(\bar{a}, \sigma_a^2)$ denotes a Gaussian random variable $\boldsymbol{a}$ with mean $\bar{a}$ and variance σ_a^2, namely, a random variable with pdf given by

$$\boxed{\mathcal{N}_{\boldsymbol{a}}(\bar{a}, \sigma_a^2) \;\equiv\; \frac{1}{\sqrt{2\pi}\,\sigma_a}\, e^{-(a-\bar{a})^2/2\sigma_a^2}} \qquad (\textbf{notation}) \tag{3.45}$$

On the other hand, if the realization for $\boldsymbol{x}$ happens to be $x = -1$, then the random variable $\boldsymbol{y}$ will be Gaussian-distributed with mean $-\frac{1}{2}$ and same variance σ_v^2:

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|x=-1) \;=\; \mathcal{N}_{\boldsymbol{y}}\left(-1/2, \sigma_v^2\right) \tag{3.46}$$

The overall pdf of $\boldsymbol{y}$ will then be given by:

$$f_{\boldsymbol{y}}(y) \;=\; p\,\mathcal{N}_{\boldsymbol{y}}\left(1/2, \sigma_v^2\right) \;+\; (1-p)\,\mathcal{N}_{\boldsymbol{y}}\left(-1/2, \sigma_v^2\right) \tag{3.47}$$

It is clear that $\boldsymbol{x}$ alters the pdf of $\boldsymbol{y}$ so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent random variables.

3.3.2 Marginal and Conditional Distributions

Given the *joint* pdf $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$ of two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ we can use this information to determine several other distributions related to the same random variables:

(a) We can determine the *marginal* pdfs corresponding to each of the variables separately, namely, the distributions $f_{\boldsymbol{x}}(x)$ and $f_{\boldsymbol{y}}(y)$. For continuous variables, these can be obtained by integrating the joint pdf over the relevant variables such as

$$f_{\boldsymbol{x}}(x) = \int_{y\in\mathcal{Y}} f_{\boldsymbol{x},\boldsymbol{y}}(x, y)dy \tag{3.48a}$$

$$f_{\boldsymbol{y}}(y) = \int_{x\in\mathcal{X}} f_{\boldsymbol{x},\boldsymbol{y}}(x, y)dx \tag{3.48b}$$

where the sets $\{\mathcal{X}, \mathcal{Y}\}$ refer to the domains over which the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are defined. The first integral removes the contribution of $\boldsymbol{y}$ while the second integral removes the contribution of $\boldsymbol{x}$. If the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ happen to be discrete and described by their joint pmf $\mathbb{P}(\boldsymbol{x}, \boldsymbol{y})$, we would determine the marginal pmfs by using sums rather than integrals:

$$\mathbb{P}(\boldsymbol{x} = x) = \sum_{y\in\mathcal{Y}} \mathbb{P}(\boldsymbol{x} = x, \boldsymbol{y} = y) \tag{3.49a}$$

$$\mathbb{P}(\boldsymbol{y} = y) = \sum_{x\in\mathcal{X}} \mathbb{P}(\boldsymbol{x} = x, \boldsymbol{y} = y) \tag{3.49b}$$

(b) We can also determine the *conditional* pdfs corresponding to each of the variables conditioned on the other variable, namely, the distributions $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ and $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$. These can be obtained by appealing to Bayes rule:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = f_{\boldsymbol{x},\boldsymbol{y}}(x, y)/f_{\boldsymbol{y}}(\boldsymbol{y}) \tag{3.50a}$$

$$f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) = f_{\boldsymbol{x},\boldsymbol{y}}(x, y)/f_{\boldsymbol{x}}(\boldsymbol{x}) \tag{3.50b}$$

In other words, the joint pdf needs to be scaled by the marginal pdfs. For discrete random variables, we would use instead:

$$\mathbb{P}(\boldsymbol{x}=x|\boldsymbol{y}=y)=\frac{\mathbb{P}(\boldsymbol{x}=x,\boldsymbol{y}=y)}{\mathbb{P}(\boldsymbol{y}=y)} \tag{3.51a}$$

$$\mathbb{P}(\boldsymbol{y}=y|\boldsymbol{x}=x)=\frac{\mathbb{P}(\boldsymbol{x}=x,\boldsymbol{y}=y)}{\mathbb{P}(\boldsymbol{x}=x)} \tag{3.51b}$$

Example 3.9 (**Law of total probability**) Consider two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with conditional pdf $f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$ and marginal pdf $f_{\boldsymbol{x}}(x)$. Using Bayes rule we have that the joint pdf factorizes as

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y)=f_{\boldsymbol{x}}(x)\,f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \tag{3.52}$$

Marginalizing over $\boldsymbol{x}$ we arrive at the useful relation, also known as the *law of total probability*:

$$\boxed{f_{\boldsymbol{y}}(y)\;=\;\int_{x\in\mathcal{X}}f_{\boldsymbol{x}}(x)f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)dx} \tag{3.53}$$

In other words, we can recover the marginal of $\boldsymbol{y}$ from knowledge of the marginal of $\boldsymbol{x}$ and the conditional of $\boldsymbol{y}$ given $\boldsymbol{x}$.

Example 3.10 (**Useful conditional relation**) Consider three discrete random variables $\{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}\}$ and note that

$$\begin{aligned}\mathbb{P}(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{C})&=\frac{\mathbb{P}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})}{\mathbb{P}(\boldsymbol{B},\boldsymbol{C})}\\&=\frac{\cancel{\mathbb{P}(\boldsymbol{B})}\,\mathbb{P}(\boldsymbol{A}|\boldsymbol{B})\,\mathbb{P}(\boldsymbol{C}|\boldsymbol{A},\boldsymbol{B})}{\cancel{\mathbb{P}(\boldsymbol{B})}\,\mathbb{P}(\boldsymbol{C}|\boldsymbol{B})}\end{aligned} \tag{3.54}$$

so that

$$\boxed{\mathbb{P}(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{C})=\frac{\mathbb{P}(\boldsymbol{A}|\boldsymbol{B})\,\mathbb{P}(\boldsymbol{C}|\boldsymbol{A},\boldsymbol{B})}{\mathbb{P}(\boldsymbol{C}|\boldsymbol{B})}} \tag{3.55}$$

Example 3.11 (**Finding marginal and conditional distributions**) Let us consider a situation involving two discrete random variables, denoted by $\boldsymbol{C}$ (cold) and $\boldsymbol{H}$ (headache). Each variable is binary and assumes the values $\{0,1\}$. For example, $\boldsymbol{C}=1$ and $\boldsymbol{H}=0$ means that the individual has a cold but does not have a headache. Likewise, the combination $\boldsymbol{C}=0$ and $\boldsymbol{H}=0$ means that the individual neither has a cold nor a headache. There are four combinations for the random variables and we describe their joint pmf in the following tabular form (the numbers in the table are for illustration purposes only and do not correspond to any actual measurements or have any medical significance):

$\boldsymbol{C}$ (cold)	$\boldsymbol{H}$ (headache)	$\mathbb{P}(\boldsymbol{C}, \boldsymbol{H})$ (joint pmf)
0	0	0.60
0	1	0.10
1	0	0.10
1	1	0.20

Observe how all entries in the last column corresponding to the joint pmf add up to one, as expected. Let us determine first the marginal pmf for the variable $\boldsymbol{C}$. For this purpose, we need to determine the *two* probabilities: $\mathbb{P}(\boldsymbol{C} = 1)$ and $\mathbb{P}(\boldsymbol{C} = 0)$. This is because the variable $\boldsymbol{C}$ can assume one of two values in $\{0, 1\}$. For the first probability, we add over H when $\boldsymbol{C} = 1$ to get

$$\begin{aligned}\mathbb{P}(\boldsymbol{C} = 1) &= \sum_{H \in \{0,1\}} \mathbb{P}(\boldsymbol{C} = 1, \boldsymbol{H} = H) \\ &= \mathbb{P}(\boldsymbol{C} = 1, \boldsymbol{H} = 1) \ + \ \mathbb{P}(\boldsymbol{C} = 1, \boldsymbol{H} = 0) \\ &= 0.20 + 0.10 \\ &= 0.3 \end{aligned} \tag{3.56}$$

Observe that we simply added the probabilities in the last two rows of the table corresponding to $\boldsymbol{C} = 1$. We repeat for $\mathbb{P}(\boldsymbol{C} = 0)$ to find that

$$\begin{aligned}\mathbb{P}(\boldsymbol{C} = 0) &= \sum_{H \in \{0,1\}} \mathbb{P}(\boldsymbol{C} = 0, \boldsymbol{H} = H) \\ &= \mathbb{P}(\boldsymbol{C} = 0, \boldsymbol{H} = 1) \ + \ \mathbb{P}(\boldsymbol{C} = 0, \boldsymbol{H} = 0) \\ &= 0.10 + 0.60 \\ &= 0.7 \end{aligned} \tag{3.57}$$

Here we added the probabilities in the first two rows of the table corresponding to $\boldsymbol{C} = 0$. Note that the two probabilities for $\boldsymbol{C}$ add up to one, as is expected for a valid pmf. In a similar manner we find that the pmf for the variable $\boldsymbol{H}$ is given by

$$\mathbb{P}(\boldsymbol{H} = 1) = 0.3, \quad \mathbb{P}(\boldsymbol{H} = 0) = 0.7 \tag{3.58}$$

Using the marginal pmfs, we can now determine conditional pmfs. For example, assume we observe that $\boldsymbol{H} = 1$ (that is, the individual has a headache), and we would like to know the likelihood that the individual has a cold too. To do so, we appeal to Bayes rule and write

$$\mathbb{P}(\boldsymbol{C} = 1 | \boldsymbol{H} = 1) = \frac{\mathbb{P}(\boldsymbol{C} = 1, \boldsymbol{H} = 1)}{\mathbb{P}(\boldsymbol{H} = 1)} \ = \ 0.2/0.3 \ = \ 2/3 \tag{3.59}$$

Accordingly, we also have

$$\mathbb{P}(\boldsymbol{C} = 0 | \boldsymbol{H} = 1) = 1/3 \tag{3.60}$$

since these two conditional probabilities need to add up to one. In a similar manner, we find that

$$\mathbb{P}(\boldsymbol{C} = 1 | \boldsymbol{H} = 0) = 1/7, \quad \mathbb{P}(\boldsymbol{C} = 0 | \boldsymbol{H} = 0) = 6/7 \tag{3.61}$$

Observe that in order to specify the conditional pmf of $\boldsymbol{C}$ given $\boldsymbol{H}$ we need to determine four probability values since each of the variables $\boldsymbol{C}$ or $\boldsymbol{H}$ assumes 2 levels in $\{0, 1\}$. We repeat similar calculations to compute the conditional pmf of $\boldsymbol{H}$ given $\boldsymbol{C}$ (with the roles of $\boldsymbol{C}$ and $\boldsymbol{H}$) reversed. We summarize these conditional calculations in tabular form:

$\boldsymbol{C}$	$\boldsymbol{H}$	$\mathbb{P}(\boldsymbol{C}\|\boldsymbol{H})$ (conditional pmf)	$\mathbb{P}(\boldsymbol{H}\|\boldsymbol{C})$ (conditional pmf)
0	0	6/7	6/7
0	1	1/3	1/7
1	0	1/7	1/3
1	1	2/3	2/3

Observe that the entries in the third column corresponding to the conditional pmf of $\boldsymbol{C}$ given $\boldsymbol{H}$ do *not* add up to one. Likewise, for the last column corresponding to the conditional pmf of $\boldsymbol{H}$ given $\boldsymbol{C}$.

3.3.3 Dependent Random Variables

We say that two continuous random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ are *independent* of each other if, and only if,:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = f_{\boldsymbol{x}}(x) \quad \text{and} \quad f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) = f_{\boldsymbol{y}}(y) \tag{3.62}$$

In other words, the pdfs of $\boldsymbol{x}$ and $\boldsymbol{y}$ are not modified by conditioning on knowledge of $\boldsymbol{y}$ or $\boldsymbol{x}$. Otherwise, the random variables are *dependent*. It follows directly from Bayes rule that dependency is equivalent to saying that the joint pdf factorizes as the product of the two marginal pdfs:

$$f_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{y}) = f_{\boldsymbol{x}}(x)\, f_{\boldsymbol{y}}(y) \tag{3.63}$$

We will employ the following notation to refer to the fact that the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent:

$$\boxed{\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}} \qquad (\textbf{independent random variables}) \tag{3.64}$$

When the variables are discrete, they will be independent of each other if, and only if,

$$\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) = \mathbb{P}(\boldsymbol{x}) \quad \text{and} \quad \mathbb{P}(\boldsymbol{y}|\boldsymbol{x}) = \mathbb{P}(\boldsymbol{y}) \tag{3.65}$$

or, equivalently,

$$\mathbb{P}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{P}(\boldsymbol{x})\, \mathbb{P}(\boldsymbol{y}) \tag{3.66}$$

The notation in the last two expressions needs some clarification. Consider the equality $\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) = \mathbb{P}(\boldsymbol{x})$, where the random variables are indicated in boldface and no specific values are listed for them. This compact relation is read as follows. Regardless of which value we observe for $\boldsymbol{y}$, and for any value of $\boldsymbol{x}$, the likelihood of observing that value for $\boldsymbol{x}$ given $\boldsymbol{y}$ will remain unchanged. We can write the relation more explicitly as follows:

$$\mathbb{P}(\boldsymbol{x} = x|\boldsymbol{y} = y) = \mathbb{P}(\boldsymbol{x} = x), \quad \text{for any } x \in \mathcal{X},\, y \in \mathcal{Y} \tag{3.67}$$

or, in terms of the joint and marginal pmfs

$$\mathbb{P}(\boldsymbol{x} = x, \boldsymbol{y} = y) = \mathbb{P}(\boldsymbol{x} = x)\, \mathbb{P}(\boldsymbol{y} = y), \quad \text{for any } x \in \mathcal{X},\, y \in \mathcal{Y} \tag{3.68}$$

It is sufficient to find one combination of values (x, y) for which the equality does not hold to conclude that $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent.

Example 3.12 (**Checking for dependency**) Let us reconsider Example 3.11 involving the variables $\boldsymbol{C}$ (cold) and $\boldsymbol{H}$ (headache). We know from the calculations in the example that:

$$\mathbb{P}(\boldsymbol{C}=1)=0.3,\ \ \mathbb{P}(\boldsymbol{C}=1|\boldsymbol{H}=1)=2/3,\ \ \mathbb{P}(\boldsymbol{C}=1|\boldsymbol{H}=0)=1/7 \tag{3.69}$$
$$\mathbb{P}(\boldsymbol{C}=0)=0.7,\ \ \mathbb{P}(\boldsymbol{C}=0|\boldsymbol{H}=1)=1/3,\ \ \mathbb{P}(\boldsymbol{C}=0|\boldsymbol{H}=0)=6/7 \tag{3.70}$$

It is clear from these values that the random variables $\boldsymbol{C}$ and $\boldsymbol{H}$ are dependent; knowledge of one variable alters the likelihood of the other variable. For instance, knowing that the individual has a headache ($\boldsymbol{H}=1$), raises the likelihood of the individual having a cold from 0.3 to 2/3.

Example 3.13 (**Dependency is not causality**) The notion of statistical dependence is *bidirectional* or *symmetric*. When $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent, this means that $\boldsymbol{x}$ depends on $\boldsymbol{y}$ *and* $\boldsymbol{y}$ depends on $\boldsymbol{x}$ in the sense that their conditional pdfs (or pmfs) satisfy (3.62) or (3.65). This also means that observing one variable alters the likelihood about the other variable. Again, from Example 3.11, the variables $\boldsymbol{C}$ and $\boldsymbol{H}$ are dependent on each other. Observing whether an individual has a cold or not, alters the likelihood of whether the same individual has a headache or not. Similarly, observing whether the individual has a headache or not, alters the likelihood of that individual having a cold or not.

The notion of dependency between random variables is sometimes confused with the notion of *causality* or *causation*. Causality implies dependency but not the other way around. In other words, statistical dependence is a necessary but not sufficient condition for causality:

$$\textbf{causality} \implies \textbf{statistical dependence} \tag{3.71}$$

One main reason why these two notions are not equivalent is because dependency is symmetric while causality is *asymmetric*. If a random variable $\boldsymbol{y}$ depends *causally* on another random variable $\boldsymbol{x}$, this means that $\boldsymbol{x}$ assuming certain values will contribute to (or have an effect on) $\boldsymbol{y}$ assuming certain values of its own. For example, if $\boldsymbol{x}$ is the random variable that indicates whether it is raining or not and $\boldsymbol{y}$ is the random variable that indicates whether the grass in the garden is wet or not, then having $\boldsymbol{x}$ assume the value $\boldsymbol{x}=1$ (raining) will cause $\boldsymbol{y}$ to assume the value $\boldsymbol{y}=1$ (wet grass):

$$\boldsymbol{x}=1 \text{ (raining)} \quad \xrightarrow{\text{causes}} \quad \boldsymbol{y}=1 \text{ (wet grass)} \tag{3.72}$$

Here, we say that raining ($\boldsymbol{x}$) influences the grass ($\boldsymbol{y}$) in a causal manner. As a result, the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ will be statistically dependent as well. This is because knowing that it has rained influences the likelihood of observing a wet grass and, conversely, knowing the state of the grass influences the likelihood of having observed rain. However, while dependency is bidirectional, the same is not true for causality: observing a wet grass does not cause the rain to fall. That is, the state of the grass variable ($\boldsymbol{y}$) does not have a cause effect on the state of the rain variable ($\boldsymbol{x}$).

One formal way to define causality is to introduce the do operator. Writing $\mathsf{do}(\boldsymbol{x}=1)$ means that we manipulate the value of the random variable $\boldsymbol{x}$ and set it to one (rather than observe it as having assumed the value one). Using this abstraction, we say that a random variable $\boldsymbol{x}$ has a *cause* effect on another random variable $\boldsymbol{y}$ if, and only if,

the following *two* conditional probability relations hold:

$$\mathbb{P}\Big(\boldsymbol{y} = y|\mathsf{do}(\boldsymbol{x} = x)\Big) \;\neq\; \mathbb{P}(\boldsymbol{y} = y) \tag{3.73a}$$

$$\mathbb{P}\Big(\boldsymbol{x} = x|\mathsf{do}(\boldsymbol{y} = y)\Big) \;=\; \mathbb{P}(\boldsymbol{x} = x) \tag{3.73b}$$

for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The first relation says that having $\boldsymbol{x}$ assume particular values will alter the distribution of $\boldsymbol{y}$, while the second relation says that the reverse effect does not hold. These two conditions highlight the asymmetric nature of causality.

Example 3.14 (**Conditional independence**) We can extend the notion of independence to conditional distributions. Given three continuous random variables $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{z}$, written as

$$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\boldsymbol{z} \tag{3.74}$$

if, and only if,

$$f_{\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z}}(x, y|z) \;=\; f_{\boldsymbol{x}|\boldsymbol{z}}(x|z)\, f_{\boldsymbol{y}|\boldsymbol{z}}(y|z) \tag{3.75}$$

That is, the conditional pdf of $\{\boldsymbol{x}, \boldsymbol{y}\}$ given $\boldsymbol{z}$ decouples into the product of the individual conditional distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$ given $\boldsymbol{z}$. For discrete random variables, the independence relation translates into requiring

$$\mathbb{P}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) \;=\; \mathbb{P}(\boldsymbol{x}|\boldsymbol{z})\, \mathbb{P}(\boldsymbol{y}|\boldsymbol{z}) \tag{3.76}$$

Conditional dependence will play a prominent role in the study of Bayesian networks and probabilistic graphical models later in out treatment. Let us illustrate the definition by means of a numerical example involving three binary random variables assuming the values $\{0, 1\}$:

$$\boldsymbol{R} = \text{indicates whether it is raining (1) or not (0)} \tag{3.77a}$$

$$\boldsymbol{A} = \text{indicates whether there has been a traffic accident (1) or not (0)} \tag{3.77b}$$

$$\boldsymbol{L} = \text{indicates whether the individual is late to work (1) or not (0)} \tag{3.77c}$$

For example, $\boldsymbol{R} = 1$, $\boldsymbol{A} = 0$, and $\boldsymbol{L} = 0$ means that it is raining, there has been no traffic accident on the road, and the individual is not late to work. Since each variable is binary, there are eight possible combinations. We describe the joint pmf in the following tabular form (the numbers in the table are for illustration purposes only):

$\boldsymbol{R}$ (rain)	$\boldsymbol{A}$ (accident)	$\boldsymbol{L}$ (late)	$\mathbb{P}(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{L})$ (joint pmf)
0	0	0	4/15
0	0	1	8/45
0	1	0	1/48
0	1	1	1/16
1	0	0	2/15
1	0	1	4/45
1	1	0	1/16
1	1	1	3/16

Assume we are able to observe whether there has been an accident on the road (i.e., we are able to know whether $\boldsymbol{A} = 1$ or $\boldsymbol{A} = 0$). Given this knowledge, we want to verify whether the random variables $\boldsymbol{R}$ and $\boldsymbol{L}$ are independent of each other. In particular, if these variables turn out to be dependent, then observing the individual arriving late to work would influence the likelihood of whether it has been raining or not.

To answer these questions, we need to determine the conditional pmfs $\mathbb{P}(\boldsymbol{R}, \boldsymbol{L}|\boldsymbol{A})$, $\mathbb{P}(\boldsymbol{R}|\boldsymbol{A})$ and $\mathbb{P}(\boldsymbol{L}|\boldsymbol{A})$, and then verify whether they satisfy the product relation:

$$\mathbb{P}(\boldsymbol{R}, \boldsymbol{L}|\boldsymbol{A}) \stackrel{?}{=} \mathbb{P}(\boldsymbol{R}|\boldsymbol{A})\,\mathbb{P}(\boldsymbol{L}|\boldsymbol{A}) \tag{3.78}$$

We start by computing the marginal pmf for the variable $\boldsymbol{A}$:

$$\begin{aligned}\mathbb{P}(\boldsymbol{A}=1) &= \sum_{R\in\{0,1\}}\sum_{L\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=R, \boldsymbol{L}=L, \boldsymbol{A}=1)\\ &= 1/48 + 1/16 + 1/16 + 3/16\\ &= 1/3\end{aligned} \tag{3.79}$$

Observe that this probability is obtained by adding the entries in the last column of the table that correspond to the rows with $\boldsymbol{A}=1$. It follows that

$$\mathbb{P}(\boldsymbol{A}=0) = 2/3 \tag{3.80}$$

Next, we determine the joint pmfs $\mathbb{P}(\boldsymbol{R}, \boldsymbol{A}=1)$ and $\mathbb{P}(\boldsymbol{L}, \boldsymbol{A}=1)$:

$$\mathbb{P}(\boldsymbol{R}=1, \boldsymbol{A}=1) = \sum_{L\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=1, \boldsymbol{L}=L, \boldsymbol{A}=1) = \frac{1}{16} + \frac{3}{16} = 1/4 \tag{3.81}$$

$$\mathbb{P}(\boldsymbol{R}=0, \boldsymbol{A}=1) = \sum_{L\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=0, \boldsymbol{L}=L, \boldsymbol{A}=1) = \frac{1}{48} + \frac{1}{16} = 1/12 \tag{3.82}$$

and

$$\mathbb{P}(\boldsymbol{L}=1, \boldsymbol{A}=1) = \sum_{R\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=R, \boldsymbol{L}=1, \boldsymbol{A}=1) = \frac{1}{16} + \frac{3}{16} = 1/4 \tag{3.83}$$

$$\mathbb{P}(\boldsymbol{L}=0, \boldsymbol{A}=1) = \sum_{R\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=R, \boldsymbol{L}=0, \boldsymbol{A}=1) = \frac{1}{48} + \frac{1}{16} = 1/12 \tag{3.84}$$

We similarly determine the joint pmfs $\mathbb{P}(\boldsymbol{R}, \boldsymbol{A}=0)$ and $\mathbb{P}(\boldsymbol{L}, \boldsymbol{A}=0)$:

$$\mathbb{P}(\boldsymbol{R}=1, \boldsymbol{A}=0) = \sum_{L\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=1, \boldsymbol{L}=L, \boldsymbol{A}=0) = \frac{2}{15} + \frac{4}{45} = 2/9 \tag{3.85}$$

$$\mathbb{P}(\boldsymbol{R}=0, \boldsymbol{A}=0) = \sum_{L\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=0, \boldsymbol{L}=L, \boldsymbol{A}=0) = \frac{4}{15} + \frac{8}{45} = 4/9 \tag{3.86}$$

and

$$\mathbb{P}(\boldsymbol{L}=1, \boldsymbol{A}=0) = \sum_{R\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=R, \boldsymbol{L}=1, \boldsymbol{A}=0) = \frac{8}{15} + \frac{4}{45} = 4/15 \tag{3.87}$$

$$\mathbb{P}(\boldsymbol{L}=0, \boldsymbol{A}=0) = \sum_{R\in\{0,1\}} \mathbb{P}(\boldsymbol{R}=R, \boldsymbol{L}=0, \boldsymbol{A}=0) = \frac{4}{15} + \frac{2}{45} = 2/5 \tag{3.88}$$

Next, appealing to Bayes rule we get

$$\mathbb{P}(\boldsymbol{R}=1|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=\frac{1/4}{1/3}=3/4 \tag{3.89a}$$

$$\mathbb{P}(\boldsymbol{R}=0|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=\frac{1/12}{1/3}=1/4 \tag{3.89b}$$

$$\mathbb{P}(\boldsymbol{R}=1|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=\frac{2/9}{2/3}=1/3 \tag{3.89c}$$

$$\mathbb{P}(\boldsymbol{R}=0|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=\frac{4/9}{2/3}=2/3 \tag{3.89d}$$

and

$$\mathbb{P}(\boldsymbol{L}=1|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{L}=1,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=\frac{1/4}{1/3}=3/4 \tag{3.90a}$$

$$\mathbb{P}(\boldsymbol{L}=0|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{L}=0,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=\frac{1/12}{1/3}=1/4 \tag{3.90b}$$

$$\mathbb{P}(\boldsymbol{L}=1|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{L}=1,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=\frac{4/15}{2/3}=2/5 \tag{3.90c}$$

$$\mathbb{P}(\boldsymbol{L}=0|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{L}=0,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=\frac{2/5}{2/3}=3/5 \tag{3.90d}$$

We still need to compute the joint conditional pmf $\mathbb{P}(\boldsymbol{R},\boldsymbol{L}|\boldsymbol{A})$. Thus, note that

$$\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=1|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=1,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=(3/16)/(1/3)=9/16 \tag{3.91a}$$

$$\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=0|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=0,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=(1/16)/(1/3)=3/16 \tag{3.91b}$$

$$\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=1|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=1,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=(1/16)/(1/3)=3/16 \tag{3.91c}$$

$$\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=0|\boldsymbol{A}=1)=\frac{\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=0,\boldsymbol{A}=1)}{\mathbb{P}(\boldsymbol{A}=1)}=(1/48)/(1/3)=1/16 \tag{3.91d}$$

and

$$\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=1|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=1,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=(4/45)/(2/3)=2/15 \tag{3.92a}$$

$$\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=0|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{R}=1,\boldsymbol{L}=0,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=(2/15)/(2/3)=1/5 \tag{3.92b}$$

$$\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=1|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=1,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=(8/45)/(2/3)=4/15 \tag{3.92c}$$

$$\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=0|\boldsymbol{A}=0)=\frac{\mathbb{P}(\boldsymbol{R}=0,\boldsymbol{L}=0,\boldsymbol{A}=0)}{\mathbb{P}(\boldsymbol{A}=0)}=(4/15)/(2/3)=2/5 \tag{3.92d}$$

We collect the results in tabular form and conclude from comparing the entries in the fourth and last columns that the variables $\boldsymbol{R}$ and $\boldsymbol{L}$ are *independent* conditioned on $\boldsymbol{A}$.

$\boldsymbol{R}$	$\boldsymbol{A}$	$\boldsymbol{L}$	$\mathbb{P}(\boldsymbol{R},\boldsymbol{L}\|\boldsymbol{A})$	$\mathbb{P}(\boldsymbol{R}\|\boldsymbol{A})$	$\mathbb{P}(\boldsymbol{L}\|\boldsymbol{A})$	$\mathbb{P}(\boldsymbol{R}\|\boldsymbol{A})\times\mathbb{P}(\boldsymbol{L}\|\boldsymbol{A})$
0	0	0	2/5	2/3	3/5	2/5
0	0	1	4/15	2/3	2/5	4/15
0	1	0	1/16	1/4	1/4	1/16
0	1	1	3/16	1/4	3/4	3/16
1	0	0	1/5	1/3	3/5	1/5
1	0	1	2/15	1/3	2/5	2/15
1	1	0	3/16	3/4	1/4	3/16
1	1	1	9/16	3/4	3/4	9/16

Example 3.15 (**Other conditional independence relations**) We list additional properties for conditionally independent random variables; we focus on discrete random variables for illustration purposes although the results are applicable to continuous random variables as well.

a) First, consider two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ that are independent given $\boldsymbol{z}$. It then holds:

$$\boxed{\boldsymbol{x}\perp\!\!\!\perp\boldsymbol{y}\,|\boldsymbol{z}\iff\mathbb{P}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})=\mathbb{P}(\boldsymbol{x}|\boldsymbol{z})} \tag{3.93}$$

Proof: Indeed, note from Bayes rule that

$$\begin{aligned}\mathbb{P}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) &= \frac{\mathbb{P}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{\mathbb{P}(\boldsymbol{y},\boldsymbol{z})}\\ &= \frac{\cancel{\mathbb{P}(\boldsymbol{z})}\,\mathbb{P}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}{\cancel{\mathbb{P}(\boldsymbol{z})}\,\mathbb{P}(\boldsymbol{y}|\boldsymbol{z})}\\ &\overset{(3.76)}{=} \frac{\mathbb{P}(\boldsymbol{x}|\boldsymbol{z})\,\mathbb{P}(\boldsymbol{y}|\boldsymbol{z})}{\mathbb{P}(\boldsymbol{y}|\boldsymbol{z})},\ \text{since } \boldsymbol{x}\perp\!\!\!\perp\boldsymbol{y}\,|\,\boldsymbol{z}\\ &= \mathbb{P}(\boldsymbol{x}|\boldsymbol{z})\end{aligned} \tag{3.94}$$

■

b) Second, we consider the following result referred to as the *weak union* property for conditional independence:

$$\boxed{\boldsymbol{x} \perp\!\!\!\perp \{\boldsymbol{y}, \boldsymbol{z}\} \implies (\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\boldsymbol{z}) \text{ and } (\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{z} \,|\boldsymbol{y})} \tag{3.95}$$

That is, if $\boldsymbol{x}$ is independent of both $\boldsymbol{y}$ and $\boldsymbol{z}$, then $\boldsymbol{x}$ is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{z}$ and of $\boldsymbol{z}$ given $\boldsymbol{y}$.

Proof: Since $\boldsymbol{x}$ is independent of both $\boldsymbol{y}$ and $\boldsymbol{z}$, it holds that

$$\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) = \mathbb{P}(\boldsymbol{x}), \quad \mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) = \mathbb{P}(\boldsymbol{x}), \quad \mathbb{P}(\boldsymbol{x}|\boldsymbol{z}) = \mathbb{P}(\boldsymbol{x}) \tag{3.96}$$

Consequently, we have

$$\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) = \mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) \quad \text{and} \quad \mathbb{P}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) = \mathbb{P}(\boldsymbol{x}|\boldsymbol{z}) \tag{3.97}$$

which, in view of (3.93) allow us to conclude that $\boldsymbol{x}$ is independent of $\boldsymbol{z}$ given $\boldsymbol{y}$ and of $\boldsymbol{y}$ given $\boldsymbol{z}$.

■

c) Third, we consider the following result referred to as the *contraction* property for conditional independence:

$$\boxed{(\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{z}) \quad \text{and} \quad (\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{z}) \implies (\boldsymbol{x} \perp\!\!\!\perp \{\boldsymbol{y}, \boldsymbol{z}\})} \tag{3.98}$$

That is, if $\boldsymbol{x}$ is independent of $\boldsymbol{z}$ and conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{z}$, then $\boldsymbol{x}$ is independent of both $\boldsymbol{y}$ and $\boldsymbol{z}$.

Proof: From the assumed independence properties we have

$$\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) = \mathbb{P}(\boldsymbol{x}|\boldsymbol{z}) = \mathbb{P}(\boldsymbol{x}) \tag{3.99}$$

from which we conclude that $\boldsymbol{x}$ is independent of both $\boldsymbol{y}$ and $\boldsymbol{z}$.

■

3.3.4 Conditional Mean

The conditional mean of a real-valued random variable $\boldsymbol{x}$ given observations of another real-valued random variable $\boldsymbol{y}$ denoted by $\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$ and is defined as the calculation:

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}) \;=\; \int_{x \in \mathcal{X}} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) dx \tag{3.100}$$

where both variables are assumed to be continuous in this representation. This computation amounts to determining the center of gravity of the conditional pdf of $\boldsymbol{x}$ given $\boldsymbol{y}$. When both variables are discrete, expression (3.100) is replaced by

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}) \;=\; \sum_{m=1}^{M} x_m \, \mathbb{P}(\boldsymbol{x} = x_m | \boldsymbol{y} = y) \tag{3.101}$$

where we are assuming that $\boldsymbol{x}$ admits M possible outcomes $\{x_m\}$ with probability p_m each. The next example considers a situation where one random variable is continuous and the other is discrete.

Example 3.16 (**Conditional mean computation**) Assume $\boldsymbol{y}$ is a random variable that is red with probability 1/3 and blue with probability 2/3:

$$\mathbb{P}(\boldsymbol{y} = \text{red}) = 1/3, \quad \mathbb{P}(\boldsymbol{y} = \text{blue}) = 2/3 \tag{3.102}$$

Likewise, assume $\boldsymbol{x}$ is a random variable that is Gaussian with mean 1 and variance 2 if $\boldsymbol{y}$ is red, and uniformly distributed between -1 and 1 if $\boldsymbol{y}$ is blue. Then, the conditional pdfs of $\boldsymbol{x}$ given observations of $\boldsymbol{y}$ are given by:

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y = \text{red}) = \mathcal{N}_{\boldsymbol{x}}(1, 2), \quad f_{\boldsymbol{x}|\boldsymbol{y}}(x|y = \text{blue}) = \mathcal{U}[-1, 1] \tag{3.103}$$

where we are using the notation $\mathcal{U}[a, b]$ to refer to a uniform distribution in the interval $[a, b]$. It follows that the conditional means of $\boldsymbol{x}$ are:

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = \text{red}) = 1, \quad \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = \text{blue}) = 0 \tag{3.104}$$

We can now employ these values, along with the pmf of the discrete variable $\boldsymbol{y}$ to compute the mean of $\boldsymbol{x}$. Thus, note that the mean of $\boldsymbol{x}$ is equal to one with probability 1/3 and to zero with probability 2/3. It follows that:

$$\begin{aligned}\mathbb{E}\,\boldsymbol{x} &= \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = \text{red}) \times \mathbb{P}(\boldsymbol{y} = \text{red}) \;+\; \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = \text{blue}) \times \mathbb{P}(\boldsymbol{y} = \text{blue}) \\ &= 1 \times \frac{1}{3} \;+\; 0 \times \frac{2}{3} \;=\; \frac{1}{3}\end{aligned} \tag{3.105}$$

An alternative way to understand this result is to introduce the variable $\boldsymbol{z} = \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$. This is a discrete random variable with two values: $\boldsymbol{z} = 1$ (which happens when $\boldsymbol{y}$ is red with probability 1/3) and $\boldsymbol{z} = 0$ (which happens when $\boldsymbol{y}$ is blue with probability 2/3). That is,

$$\mathbb{P}(\boldsymbol{z} = 1) = 1/3, \quad \mathbb{P}(\boldsymbol{z} = 0) = 2/3 \tag{3.106}$$

Now, it is shown in Prob. 3.25 that, for any two random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$, it holds that

$$\boxed{\mathbb{E}\left\{\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})\right\} = \mathbb{E}\,\boldsymbol{x}} \tag{3.107}$$

where the outermost expectation is over the pdf of $\boldsymbol{y}$ while the innermost expectation is over the conditional pdf $\boldsymbol{x}$ given $\boldsymbol{y}$. We can indicate these facts explicitly by adding subscripts and writing

$$\mathbb{E}_{\,\boldsymbol{y}}\left\{\mathbb{E}_{\,\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})\right\} = \mathbb{E}\,\boldsymbol{x} \tag{3.108}$$

In this way, the desired mean of $\boldsymbol{x}$ is simply the mean of $\boldsymbol{z}$ itself. We conclude that

$$\begin{aligned}\mathbb{E}\,\boldsymbol{x} = \mathbb{E}\,\boldsymbol{z} &= 1 \times \mathbb{P}(\boldsymbol{z} = 1) \;+\; 0 \times \mathbb{P}(\boldsymbol{z} = 0) \\ &= 1 \times \frac{1}{3} \;+\; 0 \times \frac{2}{3} \;=\; 1/3\end{aligned} \tag{3.109}$$

Example 3.17 (**Another conditional mean computation**) Assume $\boldsymbol{x}$ is a binary random variable that assumes the values ± 1 with probability 1/2 each. Assume in addition that we observe a noisy realization of $\boldsymbol{x}$, say,

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v} \tag{3.110}$$

where $\boldsymbol{v}$ is a zero-mean Gaussian random variable that is independent of $\boldsymbol{x}$ and has variance equal to one, i.e., its pdf is given by

$$f_{\boldsymbol{v}}(v) = \frac{1}{\sqrt{2\pi}}\, e^{-v^2/2} \tag{3.111}$$

Let us evaluate the conditional mean of $\boldsymbol{x}$ given observations of $\boldsymbol{y}$. From definition

(3.100), we find that we need to know the conditional pdf, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, in order to evaluate the integral expression. For this purpose, we call upon future result (3.160), which states that the pdf of the sum of two independent random variables, namely, $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, is equal to the convolution of their individual pdfs, i.e.,

$$f_{\boldsymbol{y}}(y) = \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x) f_{\boldsymbol{v}}(y-x) dx \tag{3.112}$$

In this example, we have

$$f_{\boldsymbol{x}}(x) = \frac{1}{2}\delta(x-1) \;+\; \frac{1}{2}\delta(x+1) \tag{3.113}$$

where $\delta(\cdot)$ is the Dirac-delta function, so that $f_{\boldsymbol{y}}(y)$ is given by

$$f_{\boldsymbol{y}}(y) = \frac{1}{2} f_{\boldsymbol{v}}(y+1) + \frac{1}{2} f_{\boldsymbol{v}}(y-1) \tag{3.114}$$

Moreover, in view of Bayes rule (3.42b), the joint pdf of $\{\boldsymbol{x}, \boldsymbol{y}\}$ is given by

$$\begin{aligned} f_{\boldsymbol{x},\boldsymbol{y}}(x,y) &= f_{\boldsymbol{x}}(x)\, f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \\ &= \left(\frac{1}{2}\delta(x-1) \;+\; \frac{1}{2}\delta(x+1)\right) f_{\boldsymbol{v}}(y-x) \\ &= \frac{1}{2} f_{\boldsymbol{v}}(y-1)\delta(x-1) + \frac{1}{2} f_{\boldsymbol{v}}(y+1)\delta(x+1) \end{aligned} \tag{3.115}$$

Using Bayes rule again we get

$$\begin{aligned} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) &= \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)} \\ &= \frac{f_{\boldsymbol{v}}(y-1)\delta(x-1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} + \frac{f_{\boldsymbol{v}}(y+1)\delta(x+1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} \end{aligned} \tag{3.116}$$

Substituting into expression (3.100) and integrating we obtain

$$\begin{aligned} \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}) \;&=\; \frac{f_{\boldsymbol{v}}(y-1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} - \frac{f_{\boldsymbol{v}}(y+1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} \\ &=\; \frac{1}{\left(\frac{e^{-(y+1)^2/2}}{e^{-(y-1)^2/2}}\right) + 1} - \frac{1}{\left(\frac{e^{-(y-1)^2/2}}{e^{-(y+1)^2/2}}\right) + 1} \\ &=\; \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}} \\ &\triangleq\; \tanh(y) \end{aligned} \tag{3.117}$$

In other words, the conditional mean of $\boldsymbol{x}$ given observations of $\boldsymbol{y}$ is the hyperbolic tangent function, which is shown in Fig. 3.5.

3.3.5 Correlated and Orthogonal Variables

The covariance between two random variables, $\boldsymbol{x}$ and $\boldsymbol{y}$, is denoted by the symbol σ_{xy} and is defined by either of the following equivalent expressions:

$$\boxed{\sigma_{xy} \;\triangleq\; \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y}) \;=\; \mathbb{E}\,\boldsymbol{x}\boldsymbol{y} \;-\; \bar{x}\bar{y}} \qquad (\textbf{covariance}) \tag{3.118}$$

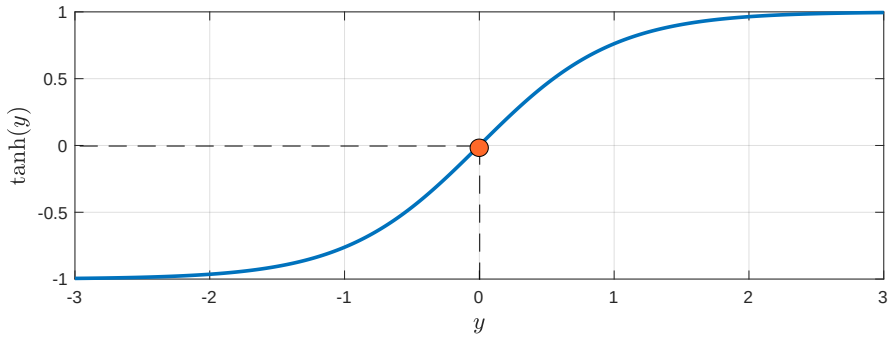

Figure 3.5 A plot of the hyperbolic tangent function, $\tanh(y) = \frac{e^{y}-e^{-y}}{e^{y}+e^{-y}}$.

We say that the random variables are uncorrelated if, and only if, their covariance is zero, i.e.,

$$\sigma_{xy} = 0 \tag{3.119}$$

which, in view of the defining relation (3.118), is also equivalent to requiring

$$\boxed{\mathbb{E}\,\boldsymbol{x}\boldsymbol{y} \;=\; (\mathbb{E}\,\boldsymbol{x})\,(\mathbb{E}\,\boldsymbol{y})} \qquad (\textbf{uncorrelated random variables}) \tag{3.120}$$

so that the mean of the product is equal to the product of the means. On the other hand, we say that two random variables are *orthogonal* if, and only if,

$$\boxed{\mathbb{E}\,\boldsymbol{x}\boldsymbol{y} = 0} \qquad (\textbf{orthogonal random variables}) \tag{3.121}$$

Observe that the means of $\boldsymbol{x}$ and $\boldsymbol{y}$ do not enter into this condition. It then follows that the concepts of orthogonality and uncorrelatedness coincide with each other if at least one of the random variables has zero mean.

When two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, it will also hold that

$$\mathbb{E}\,\boldsymbol{x}\boldsymbol{y} = (\mathbb{E}\,\boldsymbol{x})\,(\mathbb{E}\,\boldsymbol{y}) \tag{3.122}$$

This is because

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{x}\boldsymbol{y} &= \int_{x\in\mathcal{X}}\int_{y\in\mathcal{Y}} xy f_{\boldsymbol{x},\boldsymbol{y}}(x,y)dxdy \\
&\overset{(3.62)}{=} \int_{x\in\mathcal{X}}\int_{y\in\mathcal{Y}} xy f_{\boldsymbol{x}}(x) f_{\boldsymbol{y}}(y)dxdy \\
&= \left(\int_{x\in\mathcal{X}} x f_{\boldsymbol{x}}(x)dx\right)\left(\int_{y\in\mathcal{Y}} y f_{\boldsymbol{y}}(y)dy\right) \\
&= (\mathbb{E}\,\boldsymbol{x})\,(\mathbb{E}\,\boldsymbol{y})
\end{aligned} \tag{3.123}$$

It follows that independent random variables are uncorrelated:

$$\textbf{independent random variables} \implies \textbf{uncorrelated random variables} \tag{3.124}$$

The converse statement is not true.

Example 3.18 (**Uncorrelatedness and dependency**) Let $\boldsymbol{\theta}$ be a random variable that is uniformly distributed over the interval $[0, 2\pi]$. Introduce the zero-mean random variables:

$$\boldsymbol{x} = \cos\boldsymbol{\theta} \qquad \text{and} \qquad \boldsymbol{y} = \sin\boldsymbol{\theta} \tag{3.125}$$

Then, it holds that $\boldsymbol{x}^2 + \boldsymbol{y}^2 = 1$ so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent. However,

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{x}\boldsymbol{y} &= \mathbb{E}\,\cos\boldsymbol{\theta}\sin\boldsymbol{\theta} \\
&= \frac{1}{2}\mathbb{E}\,\sin 2\boldsymbol{\theta} \\
&= \frac{1}{2}\frac{1}{2\pi}\int_0^{2\pi}\sin 2\theta\, d\theta \\
&= 0
\end{aligned} \tag{3.126}$$

so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated. Therefore, we have an example of two uncorrelated random variables that are dependent.

3.4 RANDOM VECTORS

It is common in applications to encounter *vector-valued* (as opposed to scalar) random variables; also known as random vectors. A random vector consists of a collection of scalar random variables grouped together either in column form or row form. For example, assume $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two scalar random variables. Then, the column vector

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \tag{3.127a}$$

is a vector-valued random variable in column form; its dimensions are 2×1. We sometimes use the compact notation

$$\boldsymbol{x} = \text{col}\Big\{\boldsymbol{x}_1, \boldsymbol{x}_2\Big\} \tag{3.127b}$$

to denote a column vector with entries $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ stacked on top of each other. Alternatively, we could have collected the entries $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ into a row vector and obtained a row random vector instead, say,

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} \tag{3.127c}$$

In this case, the dimensions of the random vector are 1×2. Working with either the column or row format is generally a matter of convenience.

We continue, for illustration purposes, with the 2×1 vector $\boldsymbol{x} = \text{col}\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$. The mean of $\boldsymbol{x}$ is defined as the vector of individual means, namely,

$$\bar{x} \;=\; \mathbb{E}\,\boldsymbol{x} \;\triangleq\; \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} \;=\; \begin{bmatrix} \mathbb{E}\,\boldsymbol{x}_1 \\ \mathbb{E}\,\boldsymbol{x}_2 \end{bmatrix} \tag{3.128}$$

The definition extends trivially to vectors of larger dimensions so that the mean of a random vector is the vector of individual means.

With regards to the "variance" of a random vector, it will now be a matrix (and not a scalar) and will be referred to as the *covariance* matrix (rather than the variance). For the same 2×1 vector $\boldsymbol{x}$ as above, its covariance matrix is denoted by R_x and is defined as the following 2×2 matrix:

$$R_x = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1,x_2} \\ \sigma_{x_2,x_1} & \sigma_{x_2}^2 \end{bmatrix} \tag{3.129}$$

in terms of the variances $\{\sigma_{x_1}^2, \sigma_{x_2}^2\}$ of the individual entries $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$,

$$\sigma_{x_1}^2 = \mathbb{E}\,(\boldsymbol{x}_1 - \bar{x}_1)^2 \tag{3.130a}$$

$$\sigma_{x_2}^2 = \mathbb{E}\,(\boldsymbol{x}_2 - \bar{x}_2)^2 \tag{3.130b}$$

and the covariances between these individual entries:

$$\sigma_{x_1,x_2} \triangleq \mathbb{E}\,(\boldsymbol{x}_1 - \bar{x}_1)(\boldsymbol{x}_2 - \bar{x}_2) \tag{3.130c}$$

$$\sigma_{x_2,x_1} \triangleq \mathbb{E}\,(\boldsymbol{x}_2 - \bar{x}_2)(\boldsymbol{x}_1 - \bar{x}_1) = \sigma_{x_1,x_2} \tag{3.130d}$$

The matrix form (3.129) can be described in the form:

$$\boxed{R_x \triangleq \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^{\mathsf{T}}} \quad (\textbf{when } \boldsymbol{x} \textbf{ is a column vector}) \tag{3.131}$$

Expression (3.131) is general and applies to random vectors $\boldsymbol{x}$ of higher dimension than 2. It is worth noting that if $\boldsymbol{x}$ were constructed instead as a *row* (rather than a column) vector, then the covariance matrix R_x in (3.131) would instead be defined as

$$R_x \triangleq \mathbb{E}\,(\boldsymbol{x} - \bar{x})^{\mathsf{T}}(\boldsymbol{x} - \bar{x}), \quad (\textbf{when } \boldsymbol{x} \textbf{ is a row vector}) \tag{3.132}$$

with the transposed term coming first. This is because it is now the product $(\boldsymbol{x} - \bar{x})^{\mathsf{T}}(\boldsymbol{x} - \bar{x})$ that yields a matrix and leads to (3.129).

We can also extend the notion of correlations to random vectors. Thus, let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two column random vectors. Their cross-covariance matrix is denoted by R_{xy} and is defined as

$$R_{xy} \triangleq \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})^{\mathsf{T}}, \quad (\boldsymbol{x} \textbf{ and } \boldsymbol{y} \textbf{ are column vectors}) \tag{3.133a}$$

$$R_{xy} \triangleq \mathbb{E}\,(\boldsymbol{x} - \bar{x})^{\mathsf{T}}(\boldsymbol{y} - \bar{y}), \quad (\boldsymbol{x} \textbf{ and } \boldsymbol{y} \textbf{ are row vectors}) \tag{3.133b}$$

Example 3.19 (**Mean and covariance of a random vector**) Let us reconsider the setting of Example 3.8 where $\boldsymbol{x}$ assumes the values ± 1 with probability p for $+1$ and $1-p$ for -1, and $\boldsymbol{v}$ is a zero-mean Gaussian random variable with variance σ_v^2. The variables $\boldsymbol{x}$ and $\boldsymbol{v}$ are further assumed to be uncorrelated and the measurement $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \frac{1}{2}\boldsymbol{x} + \boldsymbol{v} \tag{3.134}$$

Introduce the 2×1 column vector $\boldsymbol{z}=\mathrm{col}\{\boldsymbol{x},\boldsymbol{y}\}$ and let us evaluate its mean and 2×2 covariance matrix. To begin with, using (3.23), the mean of $\boldsymbol{x}$ is given by

$$\mathbb{E}\,\boldsymbol{x}=\bar{x}\;=\;p\times 1+(1-p)\times(-1)\;=2p-1 \tag{3.135}$$

and the mean-square of $\boldsymbol{x}$ is given by

$$\mathbb{E}\,\boldsymbol{x}^2=p\times(1)^2+(1-p)\times(-1)^2\;=\;1 \tag{3.136}$$

so that the variance of $\boldsymbol{x}$ is

$$\sigma_x^2=\mathbb{E}\,\boldsymbol{x}^2\;-\;(\bar{x})^2\;=\;1-(2p-1)^2\;=\;4p(1-p) \tag{3.137}$$

Now using the measurement relation (3.134), we find that

$$\mathbb{E}\,\boldsymbol{y}=\frac{1}{2}\mathbb{E}\,\boldsymbol{x}+\mathbb{E}\,\boldsymbol{v}\;=\;\frac{1}{2}(2p-1)+0\;=\;p-\frac{1}{2} \tag{3.138}$$

and we conclude that

$$\mathbb{E}\,\boldsymbol{z}\;=\;\bar{z}\;=\;\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}\;=\;\begin{bmatrix}2p-1\\ p-\frac{1}{2}\end{bmatrix} \tag{3.139}$$

Moreover, using the assumed uncorrelatedness between $\boldsymbol{x}$ and $\boldsymbol{v}$ we further get:

$$\begin{aligned}\mathbb{E}\,\boldsymbol{y}^2&=\mathbb{E}\left(\frac{1}{2}\boldsymbol{x}+\boldsymbol{v}\right)^2\\ &=\frac{1}{4}\mathbb{E}\,\boldsymbol{x}^2\;+\;\mathbb{E}\,\boldsymbol{v}^2\;+\;\mathbb{E}\,\boldsymbol{x}\boldsymbol{v}\\ &=\frac{1}{4}\;+\;\sigma_v^2+\mathbb{E}\,\boldsymbol{x}\,\mathbb{E}\,\boldsymbol{v}\\ &=\frac{1}{4}\;+\;\sigma_v^2\;+\;0,\quad\text{since }\mathbb{E}\,\boldsymbol{v}=0\\ &=\frac{1}{4}\;+\;\sigma_v^2\end{aligned} \tag{3.140}$$

It follows that

$$\begin{aligned}\sigma_y^2&=\mathbb{E}\,\boldsymbol{y}^2-\bar{y}^2\\ &=\frac{1}{4}\;+\;\sigma_v^2-\left(p-\frac{1}{2}\right)^2\\ &=p(1-p)\;+\;\sigma_v^2\end{aligned} \tag{3.141}$$

The last expression is simply stating that

$$\sigma_y^2=\frac{1}{4}\sigma_x^2+\sigma_v^2 \tag{3.142}$$

when $\boldsymbol{x}$ and $\boldsymbol{v}$ are uncorrelated. Finally, the correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\begin{aligned}\sigma_{xy}&=\mathbb{E}\,\boldsymbol{x}\boldsymbol{y}\;-\;\bar{x}\bar{y}\\ &=\mathbb{E}\,\boldsymbol{x}\left(\frac{1}{2}\boldsymbol{x}+\boldsymbol{v}\right)\;-\;(2p-1)\left(p-\frac{1}{2}\right)\\ &=\frac{1}{2}\mathbb{E}\,\boldsymbol{x}^2\;+\;\mathbb{E}\,\boldsymbol{x}\boldsymbol{v}\;-\;(2p-1)\left(p-\frac{1}{2}\right)\\ &=\frac{1}{2}\;+\;0\;-\;(2p-1)\left(p-\frac{1}{2}\right)\\ &=2p(1-p)\end{aligned} \tag{3.143}$$

We conclude that the covariance matrix of the vector $\boldsymbol{z} = \text{col}\{\boldsymbol{x}, \boldsymbol{y}\}$ is given by

$$R_z = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 4p(1-p) & 2p(1-p) \\ 2p(1-p) & p(1-p)+\sigma_v^2 \end{bmatrix} \tag{3.144}$$

3.5 PROPERTIES OF COVARIANCE MATRICES

Covariance matrices of random vectors satisfy two important properties: **(a)** they are symmetric and **(b)** they are nonnegative-definite. Both properties can be easily verified from the definition. Indeed, using the matrix property

$$(AB)^{\mathsf{T}} = B^{\mathsf{T}} A^{\mathsf{T}} \tag{3.145}$$

for any two matrices A and B of compatible dimensions, we find that for any column random vector $\boldsymbol{x}$,

$$\begin{aligned} R_x^{\mathsf{T}} &= \Big(\mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^{\mathsf{T}}\Big)^{\mathsf{T}} \\ &= \mathbb{E}\,\Big(\underbrace{(\boldsymbol{x}-\bar{x})}_{A}\underbrace{(\boldsymbol{x}-\bar{x})^{\mathsf{T}}}_{B}\Big)^{\mathsf{T}} \\ &= \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^{\mathsf{T}}, \qquad \text{using (3.145)} \\ &= R_x \end{aligned} \tag{3.146}$$

so that

$$\boxed{R_x = R_x^{\mathsf{T}}} \tag{3.147}$$

and the covariance matrix is symmetric. It follows that covariance matrices can only have real eigenvalues.

Example 3.20 (**Eigenvalues of 2×2 covariance matrix**) Consider again Example 3.19 where we encountered the covariance matrix

$$R_z \;=\; \begin{bmatrix} 4p(1-p) & 2p(1-p) \\ 2p(1-p) & p(1-p)+\sigma_v^2 \end{bmatrix} \tag{3.148}$$

The result just established about the nature of the eigenvalues of a covariance matrix can be used to affirm that the eigenvalues of the above R_z will be real-valued no matter what the values of $p \in [0,1]$ and $\sigma_v^2 \geq 0$ are! Let us verify this fact from first principles by determining the eigenvalues of R_z. Recall that the eigenvalues $\{\lambda\}$ can be determined by solving the polynomial equation (also called the characteristic polynomial of R_z):

$$\det(\lambda I_2 - R_z) = 0 \tag{3.149}$$

in terms of the determinant of the difference $\lambda I_2 - R_z$, where I_2 is the 2×2 identity matrix. We denote the characteristic polynomial by the notation $q(\lambda)$:

$$q(\lambda) \;\triangleq\; \det(\lambda I_2 - R_z) \tag{3.150}$$

In our example we have

$$\lambda I_2 - R_z \;=\; \left[\begin{array}{cc} \lambda - 4p(1-p) & -2p(1-p) \\ -2p(1-p) & \lambda - p(1-p) - \sigma_v^2 \end{array}\right] \tag{3.151}$$

so that

$$q(\lambda) = \lambda^2 \;-\; \lambda\left(5p(1-p) + \sigma_v^2\right) \;+\; 4p(1-p)\sigma_v^2 \tag{3.152}$$

which is a quadratic polynomial in λ. The discriminant of $q(\lambda)$ is given by

$$\begin{aligned} \Delta &= \left(5p(1-p) + \sigma_v^2\right)^2 \;-\; 16p(1-p)\sigma_v^2 \\ &= 25p^2(1-p)^2 \;+\; \sigma_v^4 \;-\; 6p(1-p)\sigma_v^2 \\ &= \left(5p(1-p) - \sigma_v^2\right)^2 \;+\; 4p^2(1-p)^2 \\ &\geq 0 \end{aligned} \tag{3.153}$$

so that $q(\lambda)$ has real roots. We conclude that the eigenvalues of R_z are real for any $p \in [0,1]$ and $\sigma_v^2 \geq 0$, as expected.

Covariance matrices are also non-negative definite, i.e.,

$$\boxed{R_x \geq 0} \tag{3.154}$$

To see this, let v denote an arbitrary column vector and introduce the scalar-valued random variable

$$\boldsymbol{y} = v^{\mathsf{T}}(\boldsymbol{x} - \bar{x}) \tag{3.155}$$

Then, the variable $\boldsymbol{y}$ has zero mean and its variance is given by

$$\begin{aligned} \sigma_y^2 \;&\triangleq\; \mathbb{E}\,\boldsymbol{y}^2 \\ &=\; \mathbb{E}\,v^{\mathsf{T}}(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^{\mathsf{T}}v \\ &=\; v^{\mathsf{T}}\left(\mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^{\mathsf{T}}\right)v \\ &=\; v^{\mathsf{T}}R_x v \end{aligned} \tag{3.156}$$

But the variance of any scalar-valued random variable is always nonnegative so that that $\sigma_y^2 \geq 0$. It follows that $v^{\mathsf{T}}R_x v \geq 0$ for any v. This means that R_x is nonnegative definite, as claimed. It then follows that the eigenvalues of covariance matrices are not only real but also nonnegative.

3.6 ILLUSTRATIVE APPLICATIONS

In this section, we consider some applications of the concepts covered in the chapter in the context of convolution sums, characteristic functions, random walks, and statistical mechanics. The last two examples are meant to show how randomness is useful in modeling important physical phenomena, such as Brownian motion by particles suspended in fluids and the regulation of ion channels

in cell membranes. We will encounter some of these concepts in future sections and use them to motivate useful inference and learning methods — see, e.g., the discussion in future Sec. 66.2 on Boltzmann machines.

3.6.1 Convolution Sums

We examine first the pmf of the sum of two independent discrete random variables and show that this pdf can be obtained by convolving the individual pmfs.

Thus, let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two independent *discrete* scalar-valued random variables: $\boldsymbol{a}$ assumes the integer values $n = 0, 1, \ldots, N_a$ with probabilities $\{p_a(n)\}$ each, while $\boldsymbol{b}$ assumes the integer values $n = 0, 1, \ldots, N_b$ with probabilities $\{p_b(n)\}$ each. The probabilities $\{p_a(n), p_b(n)\}$ are zero outside the respective intervals for n. Let

$$\boldsymbol{c} \triangleq \boldsymbol{a} + \boldsymbol{b} \tag{3.157}$$

denote the sum variable. The possible realizations for $\boldsymbol{c}$ are the integer values that occur between $n = 0$ and $n = N_a + N_b$. Each possible realization n occurs with some probability $p_c(n)$ that we wish to determine. For each realization m for $\boldsymbol{a}$, the corresponding realization for $\boldsymbol{b}$ should be the value $n - m$. This means that the probability of the event $\boldsymbol{c} = n$ is given by the following expression:

$$\begin{aligned} p_c(n) &\triangleq \mathbb{P}(\boldsymbol{c} = n) \\ &= \sum_{m=-\infty}^{\infty} \mathbb{P}(\boldsymbol{a} = m, \boldsymbol{b} = n - m) \\ &= \sum_{m=-\infty}^{\infty} \mathbb{P}(\boldsymbol{a} = m)\, \mathbb{P}(\boldsymbol{b} = n - m), \quad \text{(by independence)} \\ &= \sum_{m=-\infty}^{\infty} p_a(m) p_b(n - m) \end{aligned} \tag{3.158}$$

That is,

$$\boxed{p_c(n) \;=\; p_a(n) \star p_b(n)} \qquad (\textbf{convolution sum}) \tag{3.159}$$

where the equality in the third line is due to the assumed independence of the random variables $\boldsymbol{a}$ and $\boldsymbol{b}$, and the symbol $\star$ denotes the convolution operation. We therefore observe that the sequence of probabilities $\{p_c(n)\}$ is obtained by convolving the sequences $\{p_a(n), p_b(n)\}$.

Likewise, if $\boldsymbol{x}$ and $\boldsymbol{v}$ are two independent *continuous* random variables with pdfs $f_{\boldsymbol{x}}(x)$ and $f_{\boldsymbol{v}}(v)$, respectively, and $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, then the pdf of $\boldsymbol{y}$ is given by the convolution integral — see Prob. 3.29:

$$f_{\boldsymbol{y}}(y) = \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x) f_{\boldsymbol{v}}(y - x) dx \tag{3.160}$$

3.6.2 Characteristic Functions[1]

Our second example illustrates how the continuous-time Fourier transform is applicable to the study of random variables. Thus, consider a continuous random variable $\boldsymbol{x}$ with probability density function, $f_{\boldsymbol{x}}(x)$. The characteristic function of $\boldsymbol{x}$ is denoted by $\varphi_{\boldsymbol{x}}(t)$ and is defined as the mean value of the exponential variable $e^{jt\boldsymbol{x}}$, where t is a real-valued argument. That is,

$$\varphi_{\boldsymbol{x}}(t) \triangleq \mathbb{E}\, e^{jt\boldsymbol{x}}, \quad t \in \mathbb{R} \tag{3.161}$$

or, more explicitly,

$$\varphi_{\boldsymbol{x}}(t) \triangleq \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x) e^{jtx} dx, \qquad (\textbf{characteristic function}) \tag{3.162}$$

Comparing this expression with the definition for the Fourier transform of a continuous-time signal $x(t)$, shown in the first line of Table 3.2, we see that the time-variable t is replaced by x and the frequency variable Ω is replaced by $-t$.

Table 3.2 Analogy between the Fourier transform of a continuous-time signal, $x(t)$, and the characteristic function of a random variable $\boldsymbol{x}$.

signal	transform	name	definition
$x(t)$	$X(j\Omega)$	Fourier transform	$X(j\Omega) = \int_{-\infty}^{\infty} x(t) e^{-j\Omega t} dt$
$f_{\boldsymbol{x}}(x)$	$\varphi_{\boldsymbol{x}}(t)$	characteristic transform	$\varphi_{\boldsymbol{x}}(t) = \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x) e^{jtx} dx$

A useful property of the characteristic function is that the value of its successive derivatives at $t = 0$ can be used to evaluate the moments of the random variable $\boldsymbol{x}$. Specifically, it holds that

$$\mathbb{E}\, \boldsymbol{x}^k = (-j)^k \left. \frac{d^k \varphi_{\boldsymbol{x}}(t)}{dt^k} \right|_{t=0}, \quad k = 1, 2, 3, \ldots \tag{3.163}$$

in terms of the $k-$th order derivative of $\varphi_{\boldsymbol{x}}(t)$ evaluated at $t = 0$, and where $j = \sqrt{-1}$. Moreover, in a manner similar to the inversion formula for Fourier transforms, we can recover the pdf, $f_{\boldsymbol{x}}(x)$, from knowledge of the characteristic function as follows — see Probs. 3.30–3.32:

$$f_{\boldsymbol{x}}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{\boldsymbol{x}}(t) e^{-jtx} dt \tag{3.164}$$

3.6.3 Statistical Mechanics

Our next example deals with the Boltzmann distribution, which plays a critical role in statistical mechanics; it provides a powerful tool to model complex systems

[1] This section can be skipped on a first reading.

consisting of a large number of interacting components, including interactions at the molecular level. In statistical mechanics, complex systems are modeled as having a large number of microstates. Each microstate i occurs with probability p_i and is assigned an energy level E_i. The Boltzmann distribution states that the probability that a system is at state i is proportional to $e^{-\beta E_i}$, i.e.,

$$p_i \triangleq \mathbb{P}(\text{complex system is at microstate } i) = \alpha\, e^{-\beta E_i} \tag{3.165}$$

where α describes the proportionality constant and

$$\beta = \frac{1}{k_B T} \tag{3.166}$$

which involves the temperature T measured in Kelvin and the Boltmann constant $k_B = 1.38 \times 10^{-23}$ J/K (measured in Joules/Kelvin).

Boltzmann distribution. If a complex system has N microstates, then it must hold that

$$\sum_{i=1}^{N} p_i = 1 \iff \alpha \left(\sum_{i=1}^{N} e^{-\beta E_i} \right) = 1 \tag{3.167}$$

which enables us to determine the value of α. It follows that the Boltzmann distribution (also called the Gibbs distribution) is given by

$$\boxed{\mathbb{P}(\text{complex system is at microstate } i) = \frac{e^{-\beta E_i}}{\sum_{k=1}^{N} e^{-\beta E_k}}} \tag{3.168}$$

Ion channels. Let us illustrate how these results can be used to model the behavior of ion channels, which regulate the flow of ions through biological cell membranes. A simple model for ion channels assumes that they can be either closed or open. A channel in its closed state has energy E_c and a channel in its open state has energy E_o — see Fig. 3.6. These energy levels can be measured through experimentation. Using the Boltzmann distribution (3.168), we can evaluate the probability that the channel will be open as follows:

$$\mathbb{P}(\text{channel open}) = \frac{e^{-\beta E_o}}{e^{-\beta E_o} + e^{-\beta E_c}} \tag{3.169}$$

Dividing the numerator and denominator by $e^{-\beta E_o}$, we can express the probability of an open or closed channel as

$$\mathbb{P}(\text{channel open}) = \frac{1}{1 + e^{-\beta(E_c - E_o)}} \tag{3.170a}$$

$$\mathbb{P}(\text{channel closed}) = \frac{1}{1 + e^{\beta(E_c - E_o)}} \tag{3.170b}$$

Thus, observe that the probabilities of either state depend only on the difference between the energies of the states and not on the individual values of their

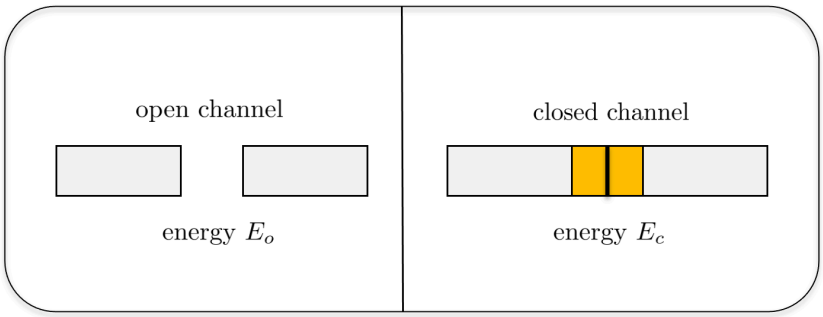

Figure 3.6 The energy of an open ion channel is assumed to be E_o and the energy of a closed ion channel is assumed to be E_c. The Boltzmann distribution allows us to evaluate the probability of encountering the channel in either state based on these energy values.

energies. Given the above probabilities, we can evaluate the average channel energy as

$$E_{\text{channel}} = \frac{E_o e^{-\beta E_o} \; + \; E_c e^{-\beta E_c}}{e^{-\beta E_o} \; + \; e^{-\beta E_c}} \tag{3.171}$$

Folding and unfolding of proteins. Similar arguments can be applied to the problem of protein folding, which refers to the process by which proteins fold into their three-dimensional structure – see Fig. 3.7. We assume the folded microstate has energy E_f and the unfolded microstate has energy E_u. We also assume that there is one folded microstate and L possible unfolded microstates. Assume we have a total of N proteins. Given this information, we would like to evaluate how many proteins on average we would encounter in their folded state among the N proteins.

The system has a total of $L+1$ microstates: one of them has energy E_f and L of them have energy E_u each. Using the Boltzmann distribution, we can evaluate the probability that the protein is folded as follows:

$$\mathbb{P}(\text{protein folded}) \; = \; \frac{e^{-\beta E_f}}{e^{-\beta E_f} \; + \; L\, e^{-\beta E_u}} \tag{3.172}$$

Dividing the numerator and the denominator by $e^{-\beta E_f}$, we can express the probability of a folded or unfolded protein as

$$\mathbb{P}(\text{protein folded}) \; = \; \frac{1}{1 \; + \; L\, e^{-\beta(E_u - E_f)}} \tag{3.173a}$$

$$\mathbb{P}(\text{protein unfolded }) \; = \; \frac{1}{1 \; + \; \frac{1}{L}\, e^{\beta(E_u - E_f)}} \tag{3.173b}$$

Observe again that the probabilities of either state depend only on the difference between the energies of the states and not on the individual values of their energies. It follows that the average number of proteins that will be encountered

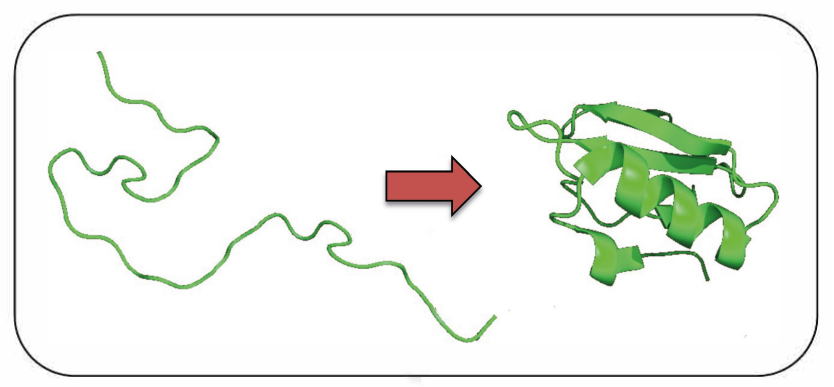

Figure 3.7 The figure shows a protein before (*left*) and after (*right*) folding into its three-dimensional structure. The source for the image is Wikimedia Commons, where the image is available in the public domain at the link https://commons.wikimedia.org/wiki/File:Protein_folding.png.

in the folded state from among a total of N proteins is:

$$N_f = N\,\mathbb{P}(\text{protein folded}) \;=\; \frac{N}{1 \;+\; L\,e^{-\beta(E_u - E_f)}} \tag{3.174}$$

3.6.4 Random Walks and Diffusion[2]

In this section we illustrate another useful application of randomness involving Brownian motion, which is used to model the random displacement of particles suspended in a fluid, such as a liquid or gas. In these environments, the random motion of particles results from collisions among molecules, leading to a random walk process where particles take successive random steps.

It is sufficient for our purposes to focus on one-dimensional (1-D) random walks; these random walks are examples of Markov chains to be discussed in greater detail in future Chapter 38. Thus, consider a particle that is initially located at the origin of displacement. At each interval of time Δt, the particle takes one step randomly along the direction of a fixed line passing through the origin (say, along the direction of the horizontal axis). The particle may step a distance d_r to the right with probability p or a distance d_ℓ to the left with probability $1-p$, as illustrated in top plot of Fig. 3.8. Let $\boldsymbol{r}$ denote the displacement of the particle after one step. Obviously, the quantity $\boldsymbol{r}$ is a random variable and we can evaluate its mean and variance. On average, the particle would be located at

$$\mathbb{E}\,\boldsymbol{r} = p\,d_r \;+\; (1-p)\,(-d_\ell) \;=\; (d_r + d_\ell)p - d_\ell \tag{3.175}$$

[2] This section can be skipped on a first reading.

and the displacement variance would be

$$\begin{aligned}\sigma_r^2 &= \mathbb{E}\,\boldsymbol{r}^2 - (\mathbb{E}\,\boldsymbol{r})^2 \\ &= pd_r^2 + (1-p)d_\ell^2 - ((d_r + d_\ell)p - d_\ell)^2 \\ &= (d_r + d_\ell)^2 p(1-p)\end{aligned} \tag{3.176}$$

In the special case when $d_r = d_\ell = \Delta x$, the above expressions simplify to

$$\mathbb{E}\,\boldsymbol{r} = \Delta x(2p-1), \quad \sigma_r^2 = 4\Delta x^2 p(1-p) \quad (\text{when } d_r = d_\ell = \Delta x) \tag{3.177}$$

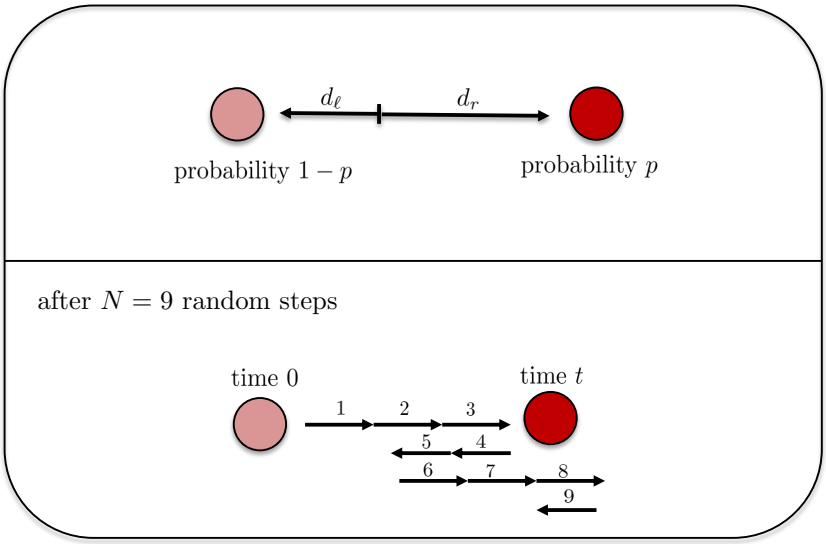

Figure 3.8 During each interval Δt, a particle takes random steps to the left (of size d_ℓ) or to the right (of size d_r) in one-dimensional space. After N such steps, and at time $t = N\Delta t$, the particle would be located at some random displacement that results from the aggregate effect of all individual steps.

After N successive independent random steps, and at time $t = N\Delta t$, the particle will be located at some displacement $\boldsymbol{x}$ along the same line. The quantity $\boldsymbol{x}$ is a random variable that is the result of summing N independent and identically distributed random variables $\{\boldsymbol{r}_n\}$ corresponding to the individual displacements over $n = 1, 2, \ldots, N$:

$$\boldsymbol{x} = \boldsymbol{r}_1 + \boldsymbol{r}_2 + \ldots + \boldsymbol{r}_N \tag{3.178}$$

Therefore, the mean and variance of $\boldsymbol{x}$ are given by

$$\mathbb{E}\,\boldsymbol{x} = N\,\mathbb{E}\,\boldsymbol{r} \;=\; N(d_r + d_\ell)p - Nd_\ell \tag{3.179}$$

$$\sigma_x^2 = N\,\sigma_r^2 \;=\; N(d_r + d_\ell)^2 p(1-p) \tag{3.180}$$

Diffusion coefficient and drift velocity. We focus henceforth on the case $d_r =$

$d_\ell = \Delta x$, where the sizes of the steps to the left or to the right are identical. Replacing N by $t/\Delta t$, we find that

$$\mathbb{E}\,\boldsymbol{x} = \frac{\Delta x}{\Delta t}\,(2p-1)\,t \tag{3.181}$$

$$\sigma_x^2 = 2\,\frac{2\Delta x^2\,p(1-p)}{\Delta t}\,t \tag{3.182}$$

The quantity

$$D \;\triangleq\; \frac{2\Delta x^2\,p(1-p)}{\Delta t}, \qquad (\textbf{diffusion coefficient}) \tag{3.183}$$

is referred to as the *diffusion coefficient* of the random walk process, while the quantity

$$v = \frac{\Delta x}{\Delta t}\,(2p-1), \quad (\textbf{drift velocity}) \tag{3.184}$$

is called the *drift velocity* of the particle. Using these variables, we find that the average displacement and variance of a random walk particle at time t are given by

$$\mathbb{E}\,\boldsymbol{x} \;=\; vt, \quad \sigma_x^2 \;=\; 2Dt \tag{3.185}$$

The variance expression can be used to estimate how far diffusing particles wander around over time. For example, assume $D = 10^{-6}\text{cm}^2/\text{sec}$ and let us estimate how far the particle wanders in 5 seconds. To do so, we evaluate the variance:

$$\sigma_x^2 = 2Dt = 10^{-5}\text{cm}^2 \tag{3.186}$$

and use the corresponding standard deviation as an estimate for the distance covered:

$$\sigma_x = \sqrt{10^{-5}} \;\approx\; 0.0032 \text{ cm} \tag{3.187}$$

Central limit theorem. When the number of steps N is large (for example, when t is large and Δt is small), then the variable $\boldsymbol{x}$ in (3.178) can be regarded as the sum of a large number of independent and identically distributed random variables. Accordingly, by the central limit theorem (see comments at end of Chapter 4), the probability density function of the displacement variable $\boldsymbol{x}$ will approach a Gaussian distribution with mean vt and variance $2Dt$, namely,

$$f_{\boldsymbol{x}}(x,t) \;=\; \frac{1}{\sqrt{4\pi Dt}}\exp\left\{-\frac{(x-vt)^2}{4Dt}\right\} \tag{3.188}$$

This distribution is a function of both x and time; it specifies the likelihood of the locations of the particle at any time t.

Einstein-Smoluchowski relation. Observe from (3.184) that the drift velocity of a diffusing particle is nonzero whenever $p \neq 1/2$. But what can cause a particle in a random-walk motion to operate under $p \neq 1/2$ and give preference to one direction of motion over another? This preferential motion can be the result of

an external force, such as gravity, which would result in a nonzero drift velocity. For example, assume a particle of mass m is diffusing down a fluid — see Fig. 3.9. Two forces act on the particle: the force of gravity, f, which acts downwards, and a drag force, ζv, which opposes the motion of the particle and acts upwards. The drag force is proportional to the velocity v through a frictional drag coefficient denoted by ζ.

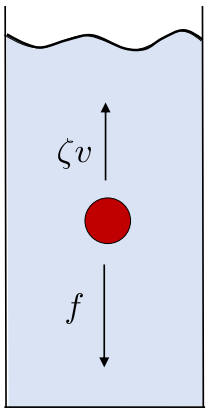

Figure 3.9 A particle of mass m diffuses down a fluid. Two forces act on it: the force of gravity, f, and the drag force, ζv.

Taking the downward direction to be the positive direction, a straightforward application of Newton's second law to the motion of the particle gives

$$m\frac{dv}{dt} \;=\; f - \zeta v \tag{3.189}$$

In steady-state, we must have $dv/dt = 0$ and, therefore, the particle attains the nonzero drift velocity $v = f/\zeta$. Substituting into (3.184), we can solve for p and deduce how the external force influences the probability value:

$$p = \frac{1}{2} \;+\; \frac{f}{2\zeta}\,\frac{\Delta x}{\Delta t} \tag{3.190}$$

It turns out that a fundamental relation exists between the diffusion coefficient, D, and the frictional coefficient, ζ, known as the Einstein-Smoluchowski relation:

$$\boxed{D\zeta \;=\; k_B T} \tag{3.191}$$

where T is the temperature in Kelvin and k_B is the Boltzmann constant. In other words, the product $D\zeta$ is constant and independent of the size of the particle and the nature of the medium where diffusion is taking place. At room temperature ($T = 300\text{K}$), the value of $k_B T$ is equal to

$$k_B T = 4.14 \times 10^{-14}\ \ \text{g}\,\text{cm}^2/\text{s}^2 \tag{3.192}$$

For a spherical particle of radius R diffusing in a medium with viscosity η ($\eta = 0.01$ g/cm s for water), the values of D and ζ are given by

$$\zeta \;=\; 6\pi\eta R, \qquad D \;=\; \frac{k_B T}{6\pi\eta R} \tag{3.193}$$

3.7 COMPLEX-VALUED VARIABLES[3]

It is common in many domains to encounter complex-valued random variables, as happens for example in the study of digital communications systems. Although the presentation in the earlier sections has focused on real-valued random variables and vectors, most of the concepts and results extend almost effortlessly to complex-valued random quantities.

A complex-valued random variable is one whose real and imaginary parts are *real*-valued random variables themselves. Specifically, if $\boldsymbol{x}$ is a scalar complex random variable, this means that it can be written in the form:

$$\boxed{\boldsymbol{x} = \boldsymbol{a} + j\boldsymbol{b}, \quad j \;\triangleq\; \sqrt{-1}} \tag{3.194}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ denote the real and imaginary parts of $\boldsymbol{x}$ and they are both real-valued random variables. Therefore, the pdf of $\boldsymbol{x}$ is completely characterized in terms of the joint pdf, $f_{\boldsymbol{a},\boldsymbol{b}}(a,b)$, of its real and imaginary parts. This means that we can regard (treat) a complex random variable as a function of two real random variables.

The mean of $\boldsymbol{x}$ is obtained as

$$\boxed{\bar{x} \;\triangleq\; \mathbb{E}\,\boldsymbol{x} \;\triangleq\; \mathbb{E}\,\boldsymbol{a} \;+\; j\mathbb{E}\,\boldsymbol{b} \;=\; \bar{a} \;+\; j\bar{b}} \tag{3.195}$$

in terms of the means of its real and imaginary parts. The variance of $\boldsymbol{x}$, on the other hand, continues to be denoted by σ_x^2 but is now defined by any of the following equivalent expressions:

$$\boxed{\sigma_x^2 \;\triangleq\; \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^* \;=\; \mathbb{E}\,|\boldsymbol{x}-\bar{x}|^2 \;=\; \mathbb{E}\,|\boldsymbol{x}|^2 \;-\; |\bar{x}|^2} \tag{3.196}$$

where the symbol $*$ denotes complex conjugation. Comparing with the earlier definition (3.15a)–(3.15b) in the real case, we see that the definition in the complex case is different because of the use of the conjugation symbol (in the real case, the conjugate of $(\boldsymbol{x}-\bar{x})$ is itself and the above definitions reduce to (3.15a)–(3.15b)). The use of the conjugate term in (3.196) is necessary in order to guarantee that σ_x^2 will remain a nonnegative number. In particular, it is immediate to verify from (3.196) that

$$\sigma_x^2 \;=\; \sigma_a^2 \;+\; \sigma_b^2 \tag{3.197}$$

[3] This section can be skipped on a first reading.

in terms of the sum of the individual variances of $\boldsymbol{a}$ and $\boldsymbol{b}$.

Likewise, the covariance between two complex-valued random variables, $\boldsymbol{x}$ and $\boldsymbol{y}$, is now defined as

$$\boxed{\sigma_{xy} \triangleq \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^*} \qquad \textbf{(covariance)} \tag{3.198}$$

with the conjugation symbol used in comparison with the real case in (3.118). We again say that the random variables are uncorrelated if, and only if, their covariance is zero, i.e.,

$$\sigma_{xy} = 0 \tag{3.199}$$

In view of the definition (3.198), this condition is equivalent to requiring

$$\boxed{\mathbb{E}\,\boldsymbol{x}\boldsymbol{y}^* = (\mathbb{E}\,\boldsymbol{x})\,(\mathbb{E}\,\boldsymbol{y})^*} \qquad \textbf{(uncorrelated random variables)} \tag{3.200}$$

On the other hand, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if, and only if,

$$\boxed{\mathbb{E}\,\boldsymbol{x}\boldsymbol{y}^* = 0} \qquad \textbf{(orthogonal random variables)} \tag{3.201}$$

It can again be verified that the concepts of orthogonality and uncorrelatedness coincide if at least one of the random variables has zero mean.

Example 3.21 **(QPSK constellation)** Consider a signal $\boldsymbol{x}$ that is chosen uniformly from a quadrature-phase-shift-keying (QPSK) constellation, i.e., $\boldsymbol{x}$ assumes any of the four values:

$$x_m = \pm\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2} \tag{3.202}$$

with equal probability $p_m = 1/4$ (see Fig. 3.10). Clearly, $\boldsymbol{x}$ is a complex-valued random variable; its mean and variance are easily found to be $\bar{x} = 0$ and $\sigma_x^2 = 1$. Indeed, note first that

$$\begin{aligned}\bar{x} &= \frac{1}{4}\left(\frac{\sqrt{2}}{2}+\frac{\sqrt{2}}{2}-\frac{\sqrt{2}}{2}-\frac{\sqrt{2}}{2}\right) + j\frac{1}{4}\left(\frac{\sqrt{2}}{2}+\frac{\sqrt{2}}{2}-\frac{\sqrt{2}}{2}-\frac{\sqrt{2}}{2}\right)\\ &= 0\end{aligned} \tag{3.203}$$

while the variance of the real part of $\boldsymbol{x} = \boldsymbol{a} + j\boldsymbol{b}$ is given by:

$$\sigma_a^2 = \frac{1}{4}\left[\left(\frac{\sqrt{2}}{2}\right)^2 + \left(\frac{\sqrt{2}}{2}\right)^2 + \left(-\frac{\sqrt{2}}{2}\right)^2 + \left(-\frac{\sqrt{2}}{2}\right)^2\right] = \frac{1}{2} \tag{3.204}$$

and, similarly, the variance of its imaginary part is $\sigma_b^2 = 1/2$. It follows that the variance of $\boldsymbol{x}$ is

$$\sigma_x^2 = \frac{1}{2} + \frac{1}{2} = 1 \tag{3.205}$$

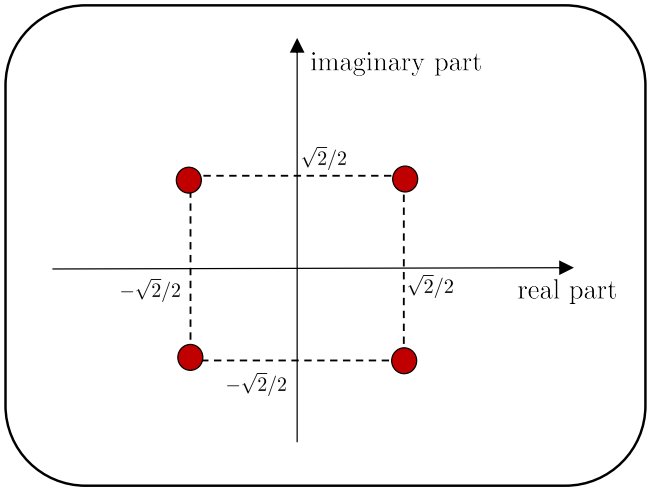

Figure 3.10 QPSK constellation with four equally probable complex symbols.

Alternatively, observe that $|\boldsymbol{x}| = 1$ for all four possibilities of $\boldsymbol{x}$, and each of these possibilities occurs with probability 1/4. Therefore,

$$\begin{aligned}\sigma_x^2 &= \mathbb{E}\,|\boldsymbol{x}|^2 \;-\; |\bar{x}|^2 \\ &= \frac{1}{4}\,(1+1+1+1) - 0 \\ &= 1\end{aligned} \tag{3.206}$$

When $\boldsymbol{x}$ is vector-valued, its mean consists of the vector of means and its covariance matrix is defined as

$$\boxed{R_x \;\triangleq\; \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^*} \qquad (\textbf{when } \boldsymbol{x} \textbf{ is a column vector}) \tag{3.207}$$

where the symbol $*$ now denotes complex-conjugate transposition (i.e., we transpose the vector and then replace each of its entries by the corresponding conjugate value). If $\boldsymbol{x}$ is instead a *row* random vector, then its covariance matrix is defined as

$$\boxed{R_x \;\triangleq\; \mathbb{E}\,(\boldsymbol{x}-\bar{x})^*(\boldsymbol{x}-\bar{x})} \qquad (\textbf{when } \boldsymbol{x} \textbf{ is a row vector}) \tag{3.208}$$

with the conjugated term coming first. This is because it is now the product $(\boldsymbol{x}-\bar{x})^*(\boldsymbol{x}-\bar{x})$ that yields a matrix.

3.8 COMMENTARIES AND DISCUSSION

Probability theory. The exposition in this chapter assumes some basic knowledge of probability theory; mainly with regards to the concepts of mean, variance, probability density function, and vector-random variables. Most of these ideas, including some of the examples, were introduced in the chapter from first principles following the overviews from Sayed (2003,2008). If additional help is needed, some accessible references on probability theory and random variables are Kolmogorov (1960), Feller (1968,1971), Billingsley (1986), Papoulis (1991), Picinbono (1993), Stark and Woods (1994), Durrett (1996), Gnedenko (1998), Chung (2000), Grimmett and Stirzaker (2001), Dudley (2002), Ash (2008), and Leon-Garcia (2008). For an insightful discussion on the notions of statistical dependence and causality, the reader may refer to Pearl (1995,2000). In Sec. 3.6.2 we illustrated how Fourier analysis is useful in the study of randomness through the notion of the characteristic function — see, e.g., Bochner (1955), Lukacs (1970), Feller (1971), and Billingsley (1986). Some good references on Fourier analysis in mathematics and signal processing are Stein and Shakarchi (2003), Katznelson (2004), Oppenheim, Schafer, and Buck (2009) and Vetterli, Kovacevic, and Goyal (2014).

The modern formulation of probability theory is due to the Soviet mathematician **Andrey Kolmogorov (1903–1987)**, who put forward in Kolmogorov (1931,1933) a collection of axioms that form the foundations for probabilistic modeling and reasoning — see the accounts by Kolmogorov (1960), Doob (1996), and Shafer and Vovk (2006). We illustrate these axioms for the case of discrete random variables.

To begin with, the finite or countable set of all possible outcomes in an experiment with discrete random results is called the *sample space*, and is denoted by the letter Ω. For example, in an experiment that involves rolling a dice, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Any subset of the sample space is called an *event*, and is denoted by the letter E. For example, observing an even outcome in the roll of the dice corresponds to observing an outcome from the event $E = \{2, 4, 6\}$. A probability measure, $\mathbb{P}(E)$, is assigned with every possible event. The three axioms of probability state that:

$$0 \leq \mathbb{P}(E) < \infty, \text{ for every } E \subset \Omega \tag{3.209a}$$

$$\mathbb{P}(\Omega) = 1 \tag{3.209b}$$

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2), \text{ for mutually exclusive events} \tag{3.209c}$$

$$\mathbb{P}\left(\bigcup_{n=1}^{N} E_n\right) = \sum_{n=1}^{N} \mathbb{P}(E_n), \text{ for mutually exclusive events} \tag{3.209d}$$

where N can be countably infinite. The first axiom states that the probability of any event is a nonnegative real number that cannot be infinite. The second axiom means that the probability of at least one event from the sample space occurring is equal to one; this statement assumes that the sample space captures all possible outcomes for the random experiment. The third equality is a special case of the last one for $N = 2$; these equalities constitute the third axiom.

Concentration inequalities. The Chebyshev inequality (3.28) is a useful result that reveals how realizations for random variables are more likely to concentrate around their means for distributions with small variances. We explain in Probs. 3.17 and 3.18 that the inequality is related to another result in probability theory known as *Markov inequality*, namely, that for any scalar nonnegative real-valued random variable, $\boldsymbol{x}$, it holds that:

$$\mathbb{P}(\boldsymbol{x} \geq \alpha) \leq \mathbb{E}\,\boldsymbol{x}/\alpha, \quad \text{for any } \alpha > 0 \tag{3.210}$$

According to Knuth (1997), the Chebyshev inequality was originally developed by Bienaymé (1853) and later proved by the Russian mathematician **Pafnuty Chebyshev**

(1821-1894) in the work by Chebyshev (1867) and subsequently by his student Markov (1884) in his PhD dissertation — see also the accounts by Hardy, Littlewood, and Pólya (1934), Bernshtein (1945), Shiryayev (1984), Papoulis (1991), and Fischer (2011).

The Markov and Chebyshev bounds are examples of *concentration inequalities*, which help bound the deviation of a random variable (or combinations of random variables) away from certain values (typically their means). In Appendix 3.B we establish three famous results known as Azuma inequality, Hoeffding inequality, and McDiarmid inequality, which provide bounds on the probability of the sum of a collection of random variables deviating from their mean. The Hoeffding inequality is due to the Finnish statistician **Wassily Hoeffding (1914-1991)** and appeared in the work by Hoeffding (1963). Earlier related investigations appear in Chernoff (1952) and Okamoto (1958). The McDiarmid inequality extends the results of Hoeffding to more general functions that satisfy a bounded variations property. This extension was proven by McDiarmid (1989). Both inequalities play an important role in the analysis of learning algorithms and will be used, for example, in the derivation of generalization bounds in future Chapter 64. For further details on concentration inequalities, readers may refer to Ledoux (2001), Boucheron, Lugosi, and Bousquet (2004), Chung and Lu (2006a,b), Massart (2007), Alon and Spencer (2008), Boucheron, Lugosi, and Massart (2013), Mohri, Rostamizadeh, and Talwalkar (2018), Vershynin (2018), and Wainwright (2019).

Bayes rule. Expression (3.39) is one manifestation of a fundamental result in probability theory known as Bayes rule, which is applicable to both cases of discrete and continuous random variables. For instance, if the letters A, B, and C denote some discrete probability events, then Bayes rule ensures that

$$\mathbb{P}(A,B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{3.211}$$

in terms of the joint probability of events A and B, and their individual and conditional probabilities. In particular, it follows that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{3.212}$$

which enables us to update the belief in event A following the observation of event B. In this way, Bayes rule allows us to update prior probabilities into posterior (conditional) probabilities. A similar construction applies when one or both random variables happen to be continuous, in which case their distributions are described in terms of probability density functions. In that situation, relation (3.211) would be replaced by (3.39), namely,

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) \;=\; f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)f_{\boldsymbol{y}}(y) \;=\; f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)f_{\boldsymbol{x}}(x) \tag{3.213}$$

A special case of Bayes rule (3.211) was first proposed by the English statistician **Thomas Bayes (1701–1761)** in his study of the problem of inferring the probability of success, p, based on observing S successes in N repeated Bernoulli trials. His work was published posthumously by the Welsh philosopher **Robert Price (1723–1791)** in the work by Bayes and Price (1763). Interestingly, Bayes rule in its general form (3.211) appears to have been independently discovered by the French mathematician **Pierre-Simon Laplace (1749–1827)** and published about a decade later in the work by Laplace (1774). The article by Stigler (1983) suggests a different historical timeline and argues that the rule may have been discovered over a decade before Bayes by another English mathematician named Nicholas Saunderson (1682–1739). However, this interpretation is not universally accepted by statisticians and the controversy remains — see, e.g., Edwards (1986), Hald (1998), Dale (2003), and Feinberg (2003).

Random walks and Brownian motion. In Sec. 3.6.4 we described one application of the central limit theorem by examining the diffusive behavior of particles. The example is motivated by the discussion in Berg (1993). The Einstein-Smoluchowski relation (3.191) is a fundamental result in physics relating the diffusion coefficient, D, of a particle and

the frictional coefficient, ζ. It was discovered independently and almost simultaneously by the German-American physicist and Nobel Laureate **Albert Einstein (1879–1955)** in the work Einstein (1905), and by Sutherland (1905) and Smoluchwski (1906) in their studies of the Brownian motion.

Brownian motion refers to the random motion of particles suspended in a fluid, where the displacements of the particles result from collisions among molecules. This explanation was provided by Einstein (1905); it was subsequently used as one indirect proof for the existence of elementary particles such as atoms and molecules. The designation "Brownian motion" is after the Scottish botanist **Robert Brown (1773–1858)** who observed under a microscope in 1827 the motion of tiny particles suspended in water — see the useful account by Pearle *et al.* (2010) and also Brown (1828,1866). There have been earlier observations of "Brownian motion" and Brown (1828) lists in his paper the names of several researchers who have commented before on aspects of this behavior. For instance, the study by van der Pas (1971) notes that the Dutch biologist **Jan Ingenhousz (1730–1799)**, who is credited with discovering the process of photosynthesis, had also reported observing the motion of coal dust particles in a liquid almost four decades before Brown in 1784 — see Ingenhousz (1784) and the English translation that appears in van der Pas (1971). In this translation, Ingenhousz comments on how *"the entire liquid and consequently everything which is contained in it, is kept in continuous motion by the evaporation, and that this motion can give the impression that some of these corpuscules are living, even if they have not the slightest life in them."*

One useful way to describe Brownian motion is in terms of a random walk process, where a particle takes successive random steps. The designation "random walk" is due to the English statistician **Karl Pearson (1857–1936)**, who is credited along with **Ronald Fisher (1890–1962)** with establishing the modern field of mathematical statistics — see the exposition by Tankard (1984). We will comment on other contributions by Pearson later in this text, including his development of the method of principal component analysis (PCA) and the Neyman-Pearson technique for hypothesis testing. For further accounts on the theory of Brownian motion and random walks, the reader may refer to several texts including by Rogers and Williams (2000), Morters and Peres (2010), Lawler and Limic (2010), Bass (2011), and Gallager (2014).

Boltzmann distribution. We described the Boltzmann distribution in expression (3.168) and explained how it is useful in characterizing the probability distribution of the states of a complex system as a function of the energies of the state levels. This probability distribution is widely used in statistical mechanics, which is the field that deals with understanding how microscopic properties at the atomic level translate into physical properties at the macroscopic level. In particular, the Boltzmann distribution encodes the useful property that lower-energy states are more likely to occur than higher-energy states — see, e.g., the treatments in Gibbs (1902), Landau and Lifshitz (1980), Hill (1987), Tolman (2010), and Pathria and Beale (2011). The distribution is named after the Austrian physicist **Ludwig Boltzmann (1844–1906)**, who is regarded as one of the developers of the field of statistical physics/mechanics. He introduced it in the work by Boltzmann (1877,1909) while developing a probabilistic view of the second law of thermodynamics. A useful historical overview of Boltzmann's work is given by Uffink (2014). Boltzmann's visionary contributions at the time, and his statistical analysis of the motion of atoms and the resulting macroscopic properties of matter, were harshly criticized by some fellow scientists who were unable to grasp his probabilistic reasoning. Unfortunately, he committed suicide in 1906. Since then, his theories and explanations have been validated by experimentation and the atomic theory of matter. We will encounter the Boltzmann distribution later in Sec. 66.2 when we study restricted Boltzmann machines in the context of deep learning networks. Applications of the Boltzmann distribution to molecular biology problems, such as the ion-channel states and protein folding examples discussed in the text, can be found in Onuchic, Luthey-Schulten, and Wolynes (1997), Huang (2005), Santana, Larranaga, and Lozano

(2008), Dubois, Gilles, and Rouzaire-Dubois (2009), and Phillips *et al.* (2012).

Law of large numbers. The weak and strong laws of large numbers are discussed in Appendix 3.A, along with various notions of convergence for random variables such as convergence in distribution, convergence in probability, almost-sure convergence, and convergence in mean-square. The weak law of large numbers is due to the Swiss mathematician **Jacob Bernoulli (1654–1705)**; it was published posthumously in his book on combinatorics by Bernoulli (1713). The law was known as the "Bernoulli theorem" for many decades and was later referred to as the "law of large numbers" by the French mathematician **Simeon Poisson (1781–1840)** in the work by Poisson (1837). The latter name has since become the common reference to these results. Many other mathematicians followed suit, refining and weakening the conditions required for the conclusions of the law to hold. In particular, the strong version of the law was first proven by the French mathematician **Emile Borel (1871–1956)** in the work by Borel (1909). Some of the weakest conditions for its validity were given later by the Russian mathematician **Andrey Kolmogorov (1903–1987)** in Kolmogorov (1927). A historical account on the laws of large numbers appears in Seneta (2013), in addition to the earlier account given by the Russian mathematician **Andrey Markov (1856–1922)**, which appears in Appendix 1 of Ondar (1981). A useful account on the strong version of the law is given by Prokhorov (2011). For technical details on the laws, the reader may consult Feller (1968,1971), Billingsley (1986,1999), Durrett (1996), and Grimmett and Stirzaker (2001).

PROBLEMS[4]

3.1 Refer to the calculations in Example 3.14 on conditional independence. Verify that the marginal pmfs for the variables $\boldsymbol{R}$ and $\boldsymbol{L}$ are given by

$$\mathbb{P}(\boldsymbol{R}=1)=17/36,\ \ \mathbb{P}(\boldsymbol{R}=0)=19/36,\ \ \mathbb{P}(\boldsymbol{L}=1)=31/60,\ \ \mathbb{P}(\boldsymbol{L}=0)=29/60$$

3.2 Refer to the calculations in Example 3.14 on conditional independence. Verify that the conditional pmf of $\boldsymbol{A}$ given $\boldsymbol{R}$ assumes the following values:

$$\begin{aligned}&\mathbb{P}(\boldsymbol{A}=1|\boldsymbol{R}=1)=9/17,\quad \mathbb{P}(\boldsymbol{A}=0|\boldsymbol{R}=1)=8/17\\ &\mathbb{P}(\boldsymbol{A}=1|\boldsymbol{R}=0)=3/19,\quad \mathbb{P}(\boldsymbol{A}=0|\boldsymbol{R}=0)=16/19\end{aligned}$$

3.3 Conclude from the results of Probs. 3.1 and 3.2 and Example 3.14 that the joint pmf of the variables $\{\boldsymbol{R},\boldsymbol{A},\boldsymbol{L}\}$ factorizes as

$$\mathbb{P}(\boldsymbol{R},\boldsymbol{A},\boldsymbol{L})=\mathbb{P}(\boldsymbol{R})\,\mathbb{P}(\boldsymbol{A}|\boldsymbol{R})\,\mathbb{P}(\boldsymbol{L}|\boldsymbol{A})$$

3.4 Refer to the calculations in Example 3.14 on conditional independence. Verify that the conditional pmf of $\boldsymbol{L}$ given $\boldsymbol{R}$ assumes the following values:

$$\begin{aligned}&\mathbb{P}(\boldsymbol{L}=1|\boldsymbol{R}=1)=199/340,\quad \mathbb{P}(\boldsymbol{L}=0|\boldsymbol{R}=1)=141/340\\ &\mathbb{P}(\boldsymbol{L}=1|\boldsymbol{R}=0)=173/380,\quad \mathbb{P}(\boldsymbol{L}=0|\boldsymbol{R}=0)=207/380\end{aligned}$$

Use the result of Prob. 3.2 to conclude that the variables $\boldsymbol{L}$ and $\boldsymbol{A}$ are not independent conditioned on $\boldsymbol{R}$ and, hence, the joint pmf of $\{\boldsymbol{R},\boldsymbol{A},\boldsymbol{L}\}$ factors in the form

$$\mathbb{P}(\boldsymbol{R},\boldsymbol{A},\boldsymbol{L})=\mathbb{P}(\boldsymbol{A})\,\mathbb{P}(\boldsymbol{R}|\boldsymbol{A})\,\mathbb{P}(\boldsymbol{L}|\boldsymbol{R},\boldsymbol{A})$$

[4] A couple of problems in this section are adapted from exercises in Sayed (2003,2008).

where the last factor cannot be replaced by $\mathbb{P}(\boldsymbol{L}|\boldsymbol{R})$. Compare with the factorization in Prob. 3.3.

3.5 Refer to the calculations in Example 3.14 on conditional independence.

Table 3.3 Joint probability mass function for the variables $\{\boldsymbol{R}, \boldsymbol{L}, \boldsymbol{A}\}$ in Prob. 3.5.

$\boldsymbol{R}$ (rain)	$\boldsymbol{A}$ (accident)	$\boldsymbol{L}$ (late)	$\mathbb{P}(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{L})$ (joint pmf)
0	0	0	0.40
0	0	1	0.05
0	1	0	0.10
0	1	1	0.10
1	0	0	0.10
1	0	1	0.10
1	1	0	0.05
1	1	1	0.10

Assume instead that the joint pmf of the variables $\{\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{L}\}$ has the values shown in Table 3.3. Repeat the derivation to verify whether the variables $\{\boldsymbol{R}, \boldsymbol{L}\}$ continue to be independent conditioned on knowledge of $\boldsymbol{A}$. Given that the individual arrived late to work, what is the likelihood that there was a traffic accident on the road?

3.6 Consider three continuous random variables $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ and assume their joint pdf factors in the form $f_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}}(x, y, z) = f_{\boldsymbol{x}}(x) f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) f_{\boldsymbol{z}|\boldsymbol{y}}(z|y)$. Verify that the variables $\{\boldsymbol{x}, \boldsymbol{z}\}$ are independent of each other conditioned on knowledge of $\boldsymbol{y}$. Verify that the same conclusion applies to discrete random variables.

3.7 Consider three discrete random variables $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ and assume their joint pmf factors in the form $\mathbb{P}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \mathbb{P}(\boldsymbol{x})\,\mathbb{P}(\boldsymbol{y})\,\mathbb{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$. Are the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ independent of each other conditioned on knowledge of $\boldsymbol{z}$? Verify that the same conclusion applies to continuous random variables.

3.8 Let $\boldsymbol{y} = \frac{1}{3}\boldsymbol{x} + \frac{1}{2}\boldsymbol{v}$, where $\boldsymbol{x}$ is uniformly distributed over the interval $[-1, 1]$ and $\boldsymbol{v}$ is a zero-mean Gaussian random variable with variance $1/2$. Both $\boldsymbol{x}$ and $\boldsymbol{v}$ are independent random variables.

(a) Find the mean and variance of $\boldsymbol{y}$.
(b) Find the correlation between $\boldsymbol{y}$ and $\boldsymbol{x}$.
(c) Find the correlation between $\boldsymbol{y}$ and $\boldsymbol{v}$.
(d) How would your answers change if $\boldsymbol{x}$ and $\boldsymbol{v}$ were only uncorrelated rather than independent?

3.9 For what values of the scalar a the matrix below is the covariance matrix of a 2×1 random vector,

$$R_z = \begin{bmatrix} 1 & a \\ a & 2 \end{bmatrix} \; ?$$

3.10 Consider a column vector $\boldsymbol{y}$ with mean $\bar{y}$ and covariance matrix R_y. What is $\mathbb{E}\,(\boldsymbol{y} \otimes \boldsymbol{y})$? Here, the symbol $\otimes$ refers to the Kronecker product operation.

3.11 If two scalar zero-mean real random variables $\boldsymbol{a}$ and $\boldsymbol{b}$ are uncorrelated, does it follow that $\boldsymbol{a}^2$ and $\boldsymbol{b}^2$ are also uncorrelated?

3.12 Consider the column vector $\boldsymbol{x} = \text{col}\{\boldsymbol{a}, \boldsymbol{b}\}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are two scalar random variables with possibly non-zero means. Use the fact that $R_x = \mathbb{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^{\mathsf{T}} \geq 0$ to establish the following Cauchy-Schwarz inequality for random variables:

$$\Big(\mathbb{E}\,(\boldsymbol{a} - \bar{a})(\boldsymbol{b} - \bar{b})\Big)^2 \;\leq\; \mathbb{E}\,(\boldsymbol{a} - \bar{a})^2 \;\times\; \mathbb{E}\,(\boldsymbol{b} - \bar{b})^2$$

3.13 Problems 3.13–3.18 are adapted from exercises in Sayed (2003,2008). Consider two scalar random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ with means $\{\bar{x}, \bar{y}\}$, variances $\{\sigma_x^2, \sigma_y^2\}$, and correlation σ_{xy}. Define the correlation coefficient $\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y$. Show that it is bounded by one, i.e., $|\rho_{xy}| \leq 1$.

3.14 A random variable $\boldsymbol{x}_1$ assumes the value $+1$ with probability p and the value -1 with probability $1-p$. A random variable $\boldsymbol{x}_2$ is distributed as follows:

$$\text{if } \boldsymbol{x}_1 = +1 \text{ then} \quad \boldsymbol{x}_2 = \begin{cases} +2 \text{ with probability } q \\ -2 \text{ with probability } 1-q \end{cases}$$

$$\text{if } \boldsymbol{x}_1 = -1 \text{ then} \quad \boldsymbol{x}_2 = \begin{cases} +3 \text{ with probability } r \\ -3 \text{ with probability } 1-r \end{cases}$$

Find the means and variances of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

3.15 Consider a Rayleigh-distributed random variable $\boldsymbol{x}$ with pdf given by (3.26). Show that its mean and variance are given by (3.27).

3.16 Suppose we observe $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{x}$ and $\boldsymbol{v}$ are independent random variables with exponential distributions with parameters λ_1 and λ_2 ($\lambda_1 \neq \lambda_2$). That is, the pdfs of $\boldsymbol{x}$ and $\boldsymbol{v}$ are $f_{\boldsymbol{x}}(x) = \lambda_1 e^{-\lambda_1 x}$ for $x \geq 0$ and $f_{\boldsymbol{v}}(v) = \lambda_2 e^{-\lambda_2 v}$ for $v \geq 0$, respectively.

(a) Using the fact that the pdf of the sum of two independent random variables is the convolution of the individual pdfs, show that

$$f_{\boldsymbol{y}}(y) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} e^{-\lambda_2 y} \left(e^{(\lambda_2 - \lambda_1) y} - 1 \right), \quad y \geq 0$$

(b) Establish that $f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = \lambda_1 \lambda_2 e^{(\lambda_2 - \lambda_1) x - \lambda_2 y}$, for $x \geq 0$ and $y \geq 0$.

3.17 Suppose $\boldsymbol{x}$ is a scalar nonnegative real-valued random variable with probability density function $f_{\boldsymbol{x}}(x)$. Show that $\mathbb{P}(\boldsymbol{x} \geq \alpha) \leq \mathbb{E}\,\boldsymbol{x}/\alpha$, for any $\alpha > 0$. This result is known as *Markov inequality*.

3.18 Consider a scalar real-valued random variable $\boldsymbol{x}$ with mean $\bar{x}$ and variance σ_x^2. Let $\boldsymbol{y} = (\boldsymbol{x} - \bar{x})^2$. Apply Markov inequality to $\boldsymbol{y}$ to establish Chebyshev inequality (3.28).

3.19 Consider a scalar real-valued random variable $\boldsymbol{x}$ with mean $\bar{x}$ and assuming values in the interval $\boldsymbol{x} \in [0, 1]$. Apply Markov inequality from Prob. 3.17 to show that, for any real number in the interval $\alpha \in (0, 1)$, it holds:

(a) $\mathbb{P}(\boldsymbol{x} > 1 - \alpha) \geq (\bar{x} - (1 - \alpha))/\alpha$.

(b) $\mathbb{P}(\boldsymbol{x} > \alpha) \geq (\bar{x} - \alpha)/(1 - \alpha)$.

3.20 Show that for any positive scalar random variable $\boldsymbol{x}$ with nonzero mean, it holds $1/\mathbb{E}\,\boldsymbol{x} < \mathbb{E}\,(1/\boldsymbol{x})$.

3.21 Suppose $\boldsymbol{x}$ is a scalar real-valued random variable with probability density function $f_{\boldsymbol{x}}(x)$ and $\mathbb{E}\,|\boldsymbol{x}|^r < \infty$. Show that, for any $\alpha > 0$ and $r > 2$, $\mathbb{P}(|\boldsymbol{x}| \geq \alpha) \leq \mathbb{E}\,|\boldsymbol{x}|^r/\alpha^r$. This result is a more general version of Markov inequality.

3.22 Consider a real-valued random variable, $\boldsymbol{x}$, with mean $\bar{x}$ and variance $\sigma_x^2 < \infty$. Let $a < b$ and $a + b = 2\bar{x}$. Conclude from Chebyshev inequality (3.28) that

$$\mathbb{P}(a < \boldsymbol{x} < b) \;\geq\; 1 - \frac{4\sigma_x^2}{(b-a)^2}$$

3.23 Consider a real-valued random variable, $\boldsymbol{x}$, with mean $\bar{x}$ and variance $\sigma_x^2 < \infty$. For any real c, conclude from the result of Prob. 3.21 that the following bound also holds:

$$\mathbb{P}(|\boldsymbol{x} - c| \geq \delta) \leq \frac{\sigma_x^2 + (\bar{x} - c)^2}{\delta^2}$$

3.24 Consider a real-valued random variable, $\boldsymbol{x}$, with mean $\bar{x}$ and variance $\sigma_x^2 < \infty$. Apply Markov inequality from Prob. 3.21 to establish the following one-sided versions of Chebyshev inequality, for any $\delta > 0$,

$$\mathbb{P}(\boldsymbol{x} \geq \bar{x} + \delta) \leq \frac{\sigma_x^2}{\sigma_x^2 + \delta^2}, \quad \mathbb{P}(\boldsymbol{x} \leq \bar{x} - \delta) \leq \frac{\sigma_x^2}{\sigma_x^2 + \delta^2}$$

3.25 Consider two real-valued random variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Establish that $\mathbb{E}\,[\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})] = \mathbb{E}\,\boldsymbol{x}$, where the outermost expectation is over the pdf of $\boldsymbol{y}$ while the innermost expectation is over the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$.

3.26 Consider two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ and a random vector $\boldsymbol{z} \in \mathbb{R}^M$ that is deterministic conditioned on knowledge of $\boldsymbol{y}$, i.e., $\mathbb{E}\,(\boldsymbol{z}|\boldsymbol{y}) = \boldsymbol{z}$. For any deterministic set $\mathcal{S} \subset \mathbb{R}^M$, establish the identity

$$\mathbb{E}\left\{\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}) \,|\, \boldsymbol{z} \in \mathcal{S}\right\} \;=\; \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{z} \in \mathcal{S})$$

where we are further conditioning on the event $\boldsymbol{z} \in \mathcal{S}$. How does this result compare to Prob. 3.25? *Remark.* This result is a special case of the law of total expectations — see, e.g., Billingsley (1986) and Weiss (2005).

3.27 Consider a discrete scalar random variable $\boldsymbol{u} = 0, 1, \ldots, N-1$, and two continuous random vector variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Assume $\boldsymbol{u}$ and $\boldsymbol{y}$ are independent of each other. Verify that

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y) = \sum_{u=0}^{N-1} \mathbb{P}(\boldsymbol{u} = u)\,\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y, \boldsymbol{u} = u)$$

3.28 The following problem is based on an exercise from Sayed (2003,2008). Consider an $M \times M$ positive-definite symmetric matrix R and introduce its eigen-decomposition, $R = \sum_{m=1}^{M} \lambda_m u_m u_m^{\mathsf{T}}$, where the λ_m are the eigenvalues of R (all positive) and the u_m are the eigenvectors of R. The u_m are orthonormal, i.e., $u_m^{\mathsf{T}} u_k = 0$ for all $m \neq k$ and $u_m^{\mathsf{T}} u_m = 1$. Let $\boldsymbol{h}$ be a random vector with probability distribution $\mathbb{P}(\boldsymbol{h} = u_m) = \lambda_m/\mathrm{Tr}(R)$, where $\mathrm{Tr}(R)$ denotes the trace of R and is equal to the sum of its eigenvalues.

(a) Show that $\mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}} = R/\mathrm{Tr}(R)$ and $\mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}}\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}} = R/\mathrm{Tr}(R)$.

(b) Show that $\mathbb{E}\,\boldsymbol{h}^{\mathsf{T}} R^{-1}\boldsymbol{h} = M/\mathrm{Tr}(R)$ and $\mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}} R^{-1}\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}} = I_M/\mathrm{Tr}(R)$, where I_M is the identity matrix of size $M \times M$.

(c) Show that $\mathbb{E}\,\boldsymbol{h}^{\mathsf{T}}\boldsymbol{h} = 1$ and $\mathbb{E}\,\boldsymbol{h} = \dfrac{1}{\mathrm{Tr}(R)} \sum_{m=1}^{M} \lambda_m u_m$.

3.29 Establish the validity of (3.160).

3.30 Starting from (3.162), verify that:

(a) $\varphi_{\boldsymbol{x}}(0) = 1$.

(b) $|\varphi_{\boldsymbol{x}}(t)| \leq 1$.

(c) $\varphi_{\boldsymbol{x}}(t) = \varphi_{\boldsymbol{x}}^{*}(-t)$.

(d) Establish the validity of (3.164).

3.31 Assume $\boldsymbol{x}$ is uniformly distributed over the interval $[a, b]$. Show that the characteristic function of $\boldsymbol{x}$ is given by

$$\varphi_{\boldsymbol{x}}(t) = \frac{e^{jtb} - e^{jta}}{jt(b-a)}$$

3.32 Assume $\boldsymbol{x}$ takes the value $x = 1$ with probability p and the value $x = 0$ with probability $1-p$. What is the characteristic function of $\boldsymbol{x}$. Use the characteristic function to evaluate all moments of $\boldsymbol{x}$.

3.33 What is the mean and variance of a Boltzmann distribution with 2 states?

3.34 If the probability of a closed ion channel is twice the probability of an open ion channel, what is the relation between the energies of the respective states?

3.35 Let $\Delta E = E_u - E_f$. Verify that the probability of encountering a folded protein can be written in the form

$$\mathbb{P}(\text{protein folded}) \;=\; \frac{1}{1 + e^{-\Delta G/k_B T}}, \quad \text{where } \Delta G \;=\; \Delta E - k_B T \ln L$$

3.36 What is the average energy of N proteins?

3.37 Let $\boldsymbol{x} = \cos\boldsymbol{\theta} + j\sin\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is uniformly distributed over the interval $[-\pi, \pi]$. Determine the mean and variance of $\boldsymbol{x}$.

3.38 Let

$$\boldsymbol{x} = \begin{bmatrix} 1+\cos\boldsymbol{\phi}+j\sin\boldsymbol{\phi} \\ \cos\boldsymbol{\phi}+j\sin\boldsymbol{\phi} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 1+\cos\boldsymbol{\theta}+j\sin\boldsymbol{\theta} \\ \cos\boldsymbol{\theta}+j\sin\boldsymbol{\theta} \end{bmatrix}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are independent of each other and uniformly distributed over the interval $[-\pi, \pi]$. Determine $\mathbb{E}\,\boldsymbol{x}$, R_x, and R_{xy}.

3.39 Conclude from the axioms of probability (3.209a)–(3.209d) that the probability of any event must be bounded by one, i.e., $\mathbb{P}(E) \leq 1$.

3.40 Conclude from the axioms of probability (3.209a)–(3.209d) that the probability of the empty event is zero, i.e., $\mathbb{P}(\emptyset) = 0$.

3.41 The pdf of a random variable $\boldsymbol{x} \geq 0$ that is exponentially-distributed with parameter $\lambda > 0$ has the form $f_{\boldsymbol{x}}(x) = \lambda e^{-\lambda x}$.

(a) Verify that $\mathbb{E}\,\boldsymbol{x} = 1/\lambda$ and $\sigma_x^2 = 1/\lambda^2$. What is the median of $\boldsymbol{x}$?

(b) Verify that the cumulative density function (cdf) of $\boldsymbol{x}$ is given by $F_{\boldsymbol{x}}(x) = 1 - e^{-\lambda x}$. Recall that the cdf at location x is defined as the area under the pdf until that location, i.e., $F_{\boldsymbol{x}}(x) = \int_{-\infty}^{x} f_{\boldsymbol{x}}(x')dx'$.

(c) Consider a sequence of positive-valued random variables $\boldsymbol{x}_n$ with cdf defined as follows:

$$F_{\boldsymbol{x}_n}(x) = 1 - (1 - \lambda/n)^{nx}, \quad x > 0$$

Show that $\boldsymbol{x}_n$ converges to the exponentially-distributed random variable $\boldsymbol{x}$ in distribution.

3.42 Consider a sequence of random variables $\boldsymbol{x}_n$ such that $\mathbb{E}\,\boldsymbol{x}_n \to \mu$ and $\sigma_{x_n}^2 \to 0$ as $n \to \infty$. Show that $\boldsymbol{x}_n \xrightarrow{p} \mu$. That is, show that $\boldsymbol{x}_n$ converges to the constant random variable μ in probability. According to definition (3.220a), this is equivalent to showing $\mathbb{P}(|\boldsymbol{x}_n - \mu| \geq \epsilon) \to 0$ for any $\epsilon \geq 0$.

3.43 Consider the random sequence $\boldsymbol{x}_n = \boldsymbol{x} + \boldsymbol{v}_n$, where the perturbation $\boldsymbol{v}_n$ has mean $\mathbb{E}\,\boldsymbol{v}_n = \mu/n^2$ and variance $\sigma_v^2 = \sigma^2/\sqrt{n}$ for some μ and $\sigma^2 > 0$. Show that the sequence $\boldsymbol{x}_n$ converges to $\boldsymbol{x}$ in probability.

3.44 Consider the random sequence $\boldsymbol{x}_n = (1 - \frac{1}{\sqrt{n}})\boldsymbol{x}$, where $\boldsymbol{x}$ is a binary random variable with $\mathbb{P}(\boldsymbol{x} = 0) = p > 0$ and $\mathbb{P}(\boldsymbol{x} = 1) = 1 - p$. Show that the sequence $\boldsymbol{x}_n$ converges to $\boldsymbol{x}$ in probability.

3.45 Consider two random sequences $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}$. Establish the following conclusions, which amount to the statement of Slutsky theorem due to Slutsky (1925) — see also Davidson (1994) and van der Vaart (2000):

(a) $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}$ and $(\boldsymbol{y}_n - \boldsymbol{x}_n) \xrightarrow{p} 0 \;\Rightarrow\; \boldsymbol{y}_n \xrightarrow{d} \boldsymbol{x}$.

(b) $\boldsymbol{x}_n \xrightarrow{p} \boldsymbol{x}$ and $(\boldsymbol{y}_n - \boldsymbol{x}_n) \xrightarrow{p} 0 \;\Rightarrow\; \boldsymbol{y}_n \xrightarrow{p} \boldsymbol{x}$.

(c) $\boldsymbol{x}_n \xrightarrow{a.s.} \boldsymbol{x}$ and $(\boldsymbol{y}_n - \boldsymbol{x}_n) \xrightarrow{a.s.} 0 \;\Rightarrow\; \boldsymbol{y}_n \xrightarrow{a.s.} \boldsymbol{x}$.

3.46 Consider a random sequence $\{\boldsymbol{x}_n\}$ that converges in distribution to $\boldsymbol{x}$, and a second random sequence $\{\boldsymbol{y}_n\}$ that converges in probability to a constant c, i.e., $\boldsymbol{x}_n \rightsquigarrow \boldsymbol{x}$ and $\boldsymbol{y}_n \xrightarrow{p} c$. Establish the following consequences of Slutsky theorem:

(a) $\boldsymbol{x}_n + \boldsymbol{y}_n \rightsquigarrow \boldsymbol{x} + c$.

(b) $\boldsymbol{x}_n \boldsymbol{y}_n \rightsquigarrow c\boldsymbol{x}$.

(c) $\boldsymbol{x}_n / \boldsymbol{y}_n \rightsquigarrow \boldsymbol{x}/c, \; c \neq 0$.

3.47 A random variable $\boldsymbol{x}$ is selected uniformly from the interval $[0, 1/2]$. Let $\boldsymbol{x}_n = 1 + 3\boldsymbol{x} + (2\boldsymbol{x})^n$.

(a) Verify that $\boldsymbol{x}_n$ approaches $1 + 3\boldsymbol{x}$ as $n \to \infty$ for any value of $\boldsymbol{x}$ in the semi-open interval $[0, 1/2)$. What happens when $\boldsymbol{x} = 1/2$?

(b) Show that the sequence $\{\boldsymbol{x}_n\}$ converges almost surely? To which random variable?

3.48 Consider a sequence of fair coin tosses with outcome $\boldsymbol{b}_n = 1$ when the coin lands a Head at the $n-$th toss or $\boldsymbol{b}_n = 0$ otherwise. We use this sequence to construct $\boldsymbol{x}_n = \prod_{m=1}^{n} \boldsymbol{b}_n$. Show that the sequence $\boldsymbol{x}_n$ converges almost surely to the constant 0.

3.49 Consider a sequence of random variables $\boldsymbol{x}_n$ that are uniformly distributed within the interval $[0 \; \frac{1}{n^2}]$. For what values of $p \geq 1$ does the sequence $\{\boldsymbol{x}_n\}$ converge in the $p-$th mean to $\boldsymbol{x} = 0$?

3.50 Consider a sequence of random variables $\{\boldsymbol{x}_n\}$ that converge in the $p-$th mean to $\boldsymbol{x}$ for some $p \geq 1$. Use the Markov inequality to conclude that the sequence $\{\boldsymbol{x}_n\}$ converges to $\boldsymbol{x}$ in probability, i.e., show that $\boldsymbol{x}_n \xrightarrow{L^p} \boldsymbol{x} \Longrightarrow \boldsymbol{x}_n \xrightarrow{p} \boldsymbol{x}$.

3.51 Show that $\boldsymbol{x}_n \xrightarrow{L^p} \boldsymbol{x} \implies \boldsymbol{x}_n \xrightarrow{L^q} \boldsymbol{x}$ for any $p > q \geq 1$. Conclude that convergence in mean-square (for which $p = 2$) implies convergence in mean (for which $p = 1$).

3.52 A biased dice is rolled once resulting in $\mathbb{P}(\text{odd}) = p$ and $\mathbb{P}(\text{even}) = 1 - p$. A sequence of random variables $\{\boldsymbol{x}_n\}$ is constructed as follows:

$$\boldsymbol{x}_n = \begin{cases} \dfrac{2n^2}{n^2 + 1/2}, & \text{when the dice roll is even} \\ 2\cos(\pi n), & \text{when the dice roll is odd} \end{cases}$$

Verify that $\mathbb{P}(\lim_{n\to\infty} \boldsymbol{x}_n = 2) = 1 - p$. Does the sequence $\{\boldsymbol{x}_n\}$ converge when the result of the dice roll is odd?

3.53 Consider the random sequence

$$\boldsymbol{x}_n = \begin{cases} \mu_1, & \text{with probability } 1 - 1/n \\ \mu_2, & \text{with probability } 1/n \end{cases}$$

(a) Show that $\boldsymbol{x}_n$ converges in the mean-square sense to $\boldsymbol{x} = \mu_1$. Does it also converge almost surely to the same limit?

(b) What happens if we change the probabilities to $1 - (1/2)^n$ and $(1/2)^n$?

3.54 Let $\{\boldsymbol{x}_n,\ n = 1, \ldots, N\}$ denote N independent scalar random variables with mean μ, with each variable satisfying $a_n \leq \boldsymbol{x}_n \leq b_n$. Let $\boldsymbol{S}_N = \sum_{n=1}^N \boldsymbol{x}_n$ denote the sum of these random variables. Let $\Delta = \sum_{n=1}^N (b_n - a_n)^2$ denote the sum of the squared lengths of the respective intervals. A famous inequality known as *Hoeffding inequality* is derived in Appendix 3.B; it asserts that for any $\delta > 0$:

$$\mathbb{P}\Big(|\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N| \geq \delta\Big) \leq 2e^{-2\delta^2/\Delta}$$

Introduce the sample average $\widehat{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n$, and assume the bounds $a_n = a$ and $b_n = b$ are uniform over n so that $a \leq \boldsymbol{x}_n \leq b$. Use Hoeffding inequality to justify the following bound, which is independent of the unknown μ:

$$\mathbb{P}\Big(|\widehat{\boldsymbol{\mu}}_N - \mu| \geq \epsilon\Big) \leq 2e^{-2\epsilon^2 N/(b-a)}$$

for any $\epsilon > 0$. Conclude the validity of the weak law of large numbers, namely, the fact that the sample average converges in probability to the actual mean as $N \to \infty$.

3.55 We continue with the setting of Prob. 3.54. Let $\{\boldsymbol{x}_n,\ n = 1, \ldots, N\}$ denote N independent scalar random variables, with each variable lying within the interval $\boldsymbol{x}_n \in [a_n, b_n]$. Introduce the sample mean:

$$\bar{\boldsymbol{x}} \triangleq \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n$$

(a) Assume $a_n = 0$ and $b_n = 1$, where all random variables lie within the interval $[0, 1]$. Use Hoeffding inequalities (3.232a)–(3.232b) to show that

$$\mathbb{P}\Big(\bar{\boldsymbol{x}} - \mathbb{E}\,\bar{\boldsymbol{x}} \geq \delta\Big) \leq e^{-2N\delta^2}, \quad \mathbb{P}\Big(|\bar{\boldsymbol{x}} - \mathbb{E}\,\bar{\boldsymbol{x}}| \geq \delta\Big) \leq 2e^{-2N\delta^2}$$

(b) Assume we wish to ensure that the likelihood (confidence level) of the sample mean $\bar{\boldsymbol{x}}$ lying within the interval $[\mathbb{E}\,\bar{\boldsymbol{x}} - \delta,\ \mathbb{E}\,\bar{\boldsymbol{x}} + \delta]$ is $1-\alpha$, for some small significance level α. Show that the number of samples needed to ensure this property is bounded by $N \leq \ln(2/\alpha)/2\delta^2$.

3.56 Consider scalar random variables $\boldsymbol{x}_n \in [0,1]$ and their zero-mean centered versions denoted by $\boldsymbol{x}_{c,n} = \boldsymbol{x}_n - \mathbb{E}\,\boldsymbol{x}_n$. Use Hoeffding lemma (3.233) to establish the following result, which provides a bound on the expectation of the maximum of a collection of centered random variables:

$$\mathbb{E}\left(\max_{1\le n\le N}\left\{\boldsymbol{x}_{c,1},\,\boldsymbol{x}_{c,2},\ldots,\,\boldsymbol{x}_{c,N}\right\}\right) \;\le\; \sqrt{\frac{1}{2}\ln N}$$

3.57 Consider two scalar random variables $\{\boldsymbol{y},\boldsymbol{z}\}$ satisfying $\mathbb{E}\,(\boldsymbol{y}|\boldsymbol{z}) = 0$. Assume there exists a function $f(\boldsymbol{z})$ and some constant $c \ge 0$ such that $f(\boldsymbol{z}) \le \boldsymbol{y} \le f(\boldsymbol{z}) + c$. Extend the derivation of the Hoeffding lemma (3.233) to verify that the following result also holds for any t:

$$\mathbb{E}\,(e^{t\boldsymbol{y}}|\boldsymbol{z}) \;\le\; e^{t^2c^2/8}$$

3.58 Consider the problem of multiplying two $N \times N$ matrices A and B to generate $C = AB$. For big data problems, the size of N can be prohibitively largely. Let $\{a_n\}$ denote the $N \times 1$ columns of A, and let $\{b_n^{\mathsf{T}}\}$ denote the $1 \times N$ rows of B. Then, C is a sum of N rank-one products of the form

$$C = \sum_{n=1}^{N} a_n b_n^{\mathsf{T}} \tag{3.214}$$

One simple approximation for computing C employs a *randomized algorithm* and is based on selecting at random R rank-one factors. Indeed, let p_n denote a discrete probability distribution over the indexes $1 \le n \le N$ with $\sum_{n=1}^{N} p_n = 1$. Select R independent integer indexes r from the range $[1, N]$ with $\mathbb{P}(r = n) = p_n$. Denote the set of selected indexes by $\mathcal{R}$ and set

$$\widehat{\boldsymbol{C}} = \sum_{r\in\mathcal{R}} \frac{1}{p_r} a_r b_r^{\mathsf{T}}$$

Verify that $\mathbb{E}\,\widehat{\boldsymbol{C}} = C$.

3.59 Consider the same setting of Prob. 3.58. We wish to select the sampling probabilities in order to minimize the mean-square-error of the difference between the approximation $\widehat{\boldsymbol{C}}$ and the product AB:

$$\{p_n^o\} \;=\; \underset{\{p_n\}}{\operatorname{argmin}}\; \mathbb{E}\,\|\widehat{\boldsymbol{C}} - AB\|_{\mathrm{F}}^2$$

Let $\|x\|$ denote the Euclidean norm of vector x. Show that the optimal probabilities are given by

$$p_n^o = \frac{\|a_n\|\,\|b_n\|}{\sum_{m=1}^{N}\|a_m\|\,\|b_m\|}$$

Remark. For additional details, the reader may refer to Drineas, Kannan, and Mahoney (2006a).

3.60 A drunken wanders randomly moving either 10m to the right or 5m to the left every minute. Where do you expect to find the drunken after 1 hour? Find the expected location and the corresponding standard deviation.

3.61 A particle wanders on average 1nm every 1 ps with velocity 1000 cm/s. What is the value of the probability p? What is the value of the diffusion coefficient?

3.62 A particle wanders on average 1nm every 1 ps with velocity -1000 cm/s (the negative sign means that the velocity is in the negative direction of the x-axis). What is the value of the probability p in this case? What is the value of the diffusion coefficient? How does it compare to the case $v = +1000$ cm/s?

3.63 What is the diffusion coefficient of a particle of radius 1nm diffusing in water?

3.64 If a particle of radius R takes t seconds to wander a distance L, how long does it take a particle of radius $R/2$ to wander for the same distance?

3.A CONVERGENCE OF RANDOM VARIABLES

There are several notions of convergence for sequences of random variables, such as convergence in probability, convergence in distribution, convergence in mean, mean-square convergence, and almost sure convergence (or convergence with probability one). We review them briefly here for ease of reference. The different notions of convergence vary in how they measure "closeness" between random variables. For additional information and proofs for some of the statements, including illustrative examples and problems, the reader may refer to Feller (1968, 1971), Billingsley (1986,1999), Davidson (1994), Durrett (1996), van der Vaart (2000), Grimmett and Stirzaker (2001), and Dudley (2002).

Convergence in distribution

Consider a sequence of scalar random variables $\{\boldsymbol{x}_n\}$ indexed by the integer n. Each variable is described by a probability density function, $f_{\boldsymbol{x}_n}(x)$. By referring to a "sequence $\{\boldsymbol{x}_n\}$" we mean that at each n, the realization for $\boldsymbol{x}_n$ arises from its pdf and the collection of these realizations will constitute the sequence. Consider further a separate random variable $\boldsymbol{x}$ with pdf $f_{\boldsymbol{x}}(x)$. The weakest notion of convergence is *convergence in distribution* (also called weak convergence). Let $F_{\boldsymbol{x}}(x)$ denote the cumulative distribution function (cdf) of the random variable $\boldsymbol{x}$; this is defined as the area under the probability density function of $\boldsymbol{x}$ up to location $\boldsymbol{x} = x$:

$$F_{\boldsymbol{x}}(x) \triangleq \int_{-\infty}^{x} f_{\boldsymbol{x}}(x')dx' = \mathbb{P}(\boldsymbol{x} \leq x) \tag{3.215}$$

Similarly, let $F_{\boldsymbol{x}_n}(x)$ denote the cdf for each $\boldsymbol{x}_n$. The sequence $\{\boldsymbol{x}_n\}$ is said to converge in distribution to the random variable $\boldsymbol{x}$ if the respective cdfs approach each other for large n at all points where $F_{\boldsymbol{x}}(x)$ is continuous, i.e.,

(convergence in distribution I)

$$\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x} \iff \lim_{n\to\infty} F_{\boldsymbol{x}_n}(x) = F_{\boldsymbol{x}}(x), \quad \{\forall\, x \,|\, F_{\boldsymbol{x}}(x) \text{ continuous}\} \tag{3.216}$$

Convergence in distribution is also denoted by the notation $\boldsymbol{x}_n \rightsquigarrow \boldsymbol{x}$. When convergence occurs, then the cdf $F_{\boldsymbol{x}}(x)$ is uniquely defined. This type of convergence depends only on the cdfs (not the actual values of the random variables) and it ensures that, for large n, the likelihoods that $\boldsymbol{x}_n$ and $\boldsymbol{x}$ lie within the same interval are essentially the same:

$$\mathbb{P}(a \leq \boldsymbol{x}_n \leq b) \approx \mathbb{P}(a \leq \boldsymbol{x} \leq b), \quad \text{for large } n \tag{3.217}$$

It also follows from $\boldsymbol{x}_n \rightsquigarrow \boldsymbol{x}$ that, for any continuous function $g(x)$, the sequence $g(\boldsymbol{x}_n)$ converges in distribution to the random variable $g(\boldsymbol{x})$. This result is known as the *continuous mapping theorem*. There are several other equivalent characterizations of convergence in distribution. One useful characterization motivated by the above remark is the following:

(convergence in distribution II)

$$\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x} \iff \lim_{n\to\infty} \mathbb{P}(\boldsymbol{x}_n \leq x) = \mathbb{P}(\boldsymbol{x} \leq x) \quad \{\forall\, x \,|\, \mathbb{P}(\boldsymbol{x} \leq x) \text{ continuous}\} \tag{3.218}$$

A second statement is that

(convergence in distribution III)

$$\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x} \iff \lim_{n\to\infty} \mathbb{E}\, g(\boldsymbol{x}_n) = \mathbb{E}\, g(\boldsymbol{x}) \tag{3.219}$$
$$\forall\, g(x)\text{: bounded and continuous or Lipschitz}$$

where a Lipschitz function is one for which $|g(a) - g(b)| \leq \delta|a - b|$ for all a, b and for some $\delta > 0$. The central limit theorem discussed later in (4.159) is one of the most famous and useful results on convergence in distribution. It is important to note though that convergence in distribution *does not* generally imply convergence of the respective probability density functions (i.e., $f_{\boldsymbol{x}_n}(x)$ need not converge to $f_{\boldsymbol{x}}(x)$). Counter examples can be found to this effect.

Convergence in probability

The second notion we consider is *convergence in probability*, which implies convergence in distribution (the converse is true only when $\boldsymbol{x}$ is the constant random variable). The sequence $\{\boldsymbol{x}_n\}$ is said to converge in probability to the random variable $\boldsymbol{x}$ if there is a high probability that the distance $|\boldsymbol{x}_n - \boldsymbol{x}|$ becomes very small for large n, i.e.,

(convergence in probability)

$$\boldsymbol{x}_n \xrightarrow{p} \boldsymbol{x} \iff \lim_{n\to\infty} \mathbb{P}(|\boldsymbol{x}_n - \boldsymbol{x}| \geq \epsilon) = 0, \quad \text{for any } \epsilon > 0 \tag{3.220a}$$
$$\iff \lim_{n\to\infty} \mathbb{P}(|\boldsymbol{x}_n - \boldsymbol{x}| < \epsilon) = 1 \tag{3.220b}$$

This definition is essentially dealing with the convergence of a sequence of probabilities. Note that checking the condition on the right-hand side requires knowledge of the joint distribution of the variables $\{\boldsymbol{x}_n, \boldsymbol{x}\}$. For random vectors $\{\boldsymbol{x}_n\}$, we would simply replace $|\boldsymbol{x}_n - \boldsymbol{x}|$ in terms of the Euclidean distance $\|\boldsymbol{x}_n - \boldsymbol{x}\|$. Interestingly, although convergence in probability does not imply the stronger notion of almost-sure convergence defined below, it can be shown that convergence in probability of a sequence $\{\boldsymbol{x}_n\}$ to $\boldsymbol{x}$ implies the existence of a subsequence $\{\boldsymbol{x}_{k_n}\}$ that converges almost surely to $\boldsymbol{x}$. The above notion of convergence in probability ensures that, in the limit, $\boldsymbol{x}_n$ will lie with high probability within the disc centered at $\boldsymbol{x}$ and of radius ϵ. The result still does not guarantee "point-wise" convergence of $\boldsymbol{x}_n$ to $\boldsymbol{x}$. This is what almost-sure convergence does.

Almost-sure convergence

The third and strongest notion we consider is *almost-sure convergence*; it implies the other two notions — see Fig. 3.11. The sequence $\{\boldsymbol{x}_n\}$ is said to converge almost surely (or with probability one) to the random variable $\boldsymbol{x}$ if there is a high probability that $\boldsymbol{x}_n$ approaches $\boldsymbol{x}$ for large n, i.e.,

(almost-sure convergence)

$$\boldsymbol{x}_n \xrightarrow{a.s.} \boldsymbol{x} \iff \mathbb{P}\Big(\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}\Big) = 1 \tag{3.221}$$

This statement guarantees the convergence of $\boldsymbol{x}_n$ to $\boldsymbol{x}$ except possibly over a set of "measure zero" — see, e.g., Prob. 3.47. In this problem, we construct a particular sequence $\boldsymbol{x}_n$ and define a separate random variable $\boldsymbol{x}$ that is uniformly distributed over the interval $[0, \frac{1}{2}]$. We then verify that $\boldsymbol{x}_n$ converges to $\boldsymbol{x}$ for all points in the semi-open interval $[0, \frac{1}{2})$ but not at the location $x = 1/2$. Since this is a singleton and the probability of $\boldsymbol{x}$ assuming values in the interval $[0, \frac{1}{2})$ is still equal to one, we are able to conclude that $\boldsymbol{x}_n$ converges almost surely to $\boldsymbol{x}$. This example clarifies the reason for the qualification "almost-surely" since some points in the domain of $\boldsymbol{x}$ may be excluded (i.e., convergence occurs for almost all points). It is not always straightforward to check

for almost-sure convergence by applying the definition (3.221). One useful *sufficient* condition is to verify that for any $\epsilon > 0$:

$$\sum_{n=1}^{\infty} \mathbb{P}(|\boldsymbol{x}_n - \boldsymbol{x}| > \epsilon) < \infty \implies \text{almost-sure convergence} \tag{3.222}$$

The *strong law of large numbers* is one of the most famous results illustrating almost-sure convergence. Consider a collection of independent and identically distributed random variables $\{\boldsymbol{x}_n\}$ with mean μ each and bounded absolute first-order moment, $\mathbb{E}\,|\boldsymbol{x}_n| < \infty$. The strong law states that the sample average estimator defined by

$$\widehat{\boldsymbol{\mu}}_N \triangleq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \tag{3.223}$$

converges almost surely to the true mean μ as $N \to \infty$, i.e.,

(strong law of large numbers)

$$\widehat{\boldsymbol{\mu}}_N \xrightarrow{a.s.} \mu \iff \mathbb{P}\Big(\lim_{N\to\infty} \widehat{\boldsymbol{\mu}}_N = \mu\Big) = 1 \tag{3.224}$$

In other words, as the number of samples increases, the likelihood that $\widehat{\boldsymbol{\mu}}_N$ will converge to the true value μ tends to one. In addition, it is possible to specify the rate of convergence of $\widehat{\boldsymbol{\mu}}_N$ towards μ as $N \to \infty$. If the variables $\boldsymbol{x}_n$ have uniform and finite variance, $\mathbb{E}\,(\boldsymbol{x}_n - \mu)^2 = \sigma_x^2 < \infty$, then it is further known that (see, e.g., Durrett (1996, p. 437)):

$$\limsup_{N\to\infty} \left\{ \frac{\widehat{\boldsymbol{\mu}}_N - \mu}{\sigma_x} \times \frac{N^{1/2}}{(2\ln\ln N)^{1/2}} \right\} = 1, \quad \text{almost surely} \tag{3.225}$$

which indicates that, for large enough N, the difference between $\widehat{\boldsymbol{\mu}}_N$ and μ is on the order of:

$$\widehat{\boldsymbol{\mu}}_N - \mu = O\left(\sqrt{\frac{2\sigma_x^2 \ln\ln N}{N}}\right) \tag{3.226}$$

using the Big-O notation. This notation will be used frequently in our treatment to compare the asymptotic convergence rates of two sequences. Thus, writing $a_n = O(b_n)$, for two sequences $\{a_n, b_n,\, n \geq 0\}$ with b_n having positive entries, means that there exists some constant $c > 0$ and index n_o such that $|a_n| \leq c b_n$ for all $n > n_o$. This also means that the decay rate of the sequence a_n is at least as fast or faster than b_n. For example, writing $a_n = O(1/n)$ means that the samples of the sequence a_n decay asymptotically at a rate that is comparable to or faster than $1/n$.

The weak version of the law of large numbers is studied in Prob. 3.54; it only ensures convergence in probability, namely, for any $\epsilon > 0$:

(weak law of large numbers)

$$\widehat{\boldsymbol{\mu}}_N \xrightarrow{p} \mu \iff \lim_{N\to\infty} \mathbb{P}\Big(|\widehat{\boldsymbol{\mu}}_N - \mu| \geq \epsilon\Big) = 0 \tag{3.227}$$

Remark 3.1. (Inference problems) The weak and strong laws of large numbers provide one useful example of how sequences of random variables arise in inference problems. In future chapters, we will describe methods that construct estimators, say, $\widehat{\boldsymbol{\theta}}_n$, for some unknown parameter $\boldsymbol{\theta}$, where n denotes the number of data points used to determine $\widehat{\boldsymbol{\theta}}_n$. We will then be interested in analyzing how well the successive estimators $\widehat{\boldsymbol{\theta}}_n$ approach $\boldsymbol{\theta}$ for increased sample sizes n. In these studies, the notions of convergence in the mean, mean-square, in distribution, and in probability will be very helpful.

■

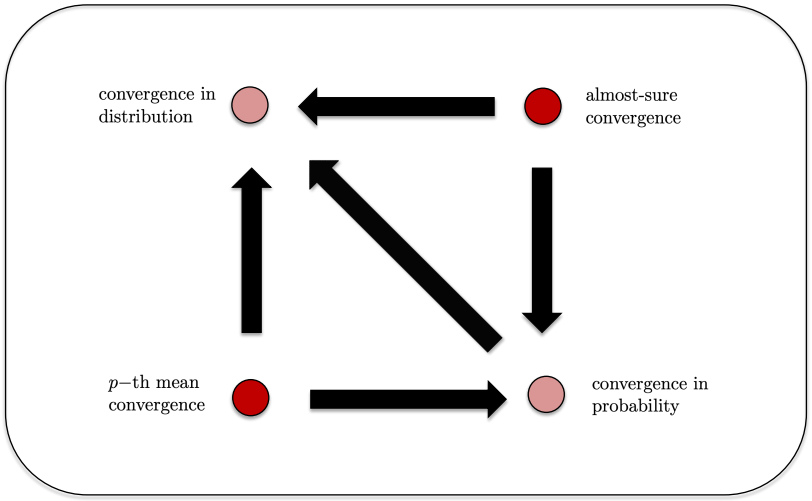

Figure 3.11 Both notions of almost-sure and $p-$th mean convergence are stronger than convergence in probability, which in turn is stronger than convergence in distribution. The direction of an arrow from location A to B means that notion A implies notion B.

Convergence in the $p-$th mean

The last and also strong notion of convergence we consider is convergence in the $p-$th mean; it is stronger than convergence in probability and convergence in distribution — see Fig. 3.11. Consider an exponent $p \geq 1$ and assume the $p-$th moments of the variables $\boldsymbol{x}_n$ are bounded, i.e., $\mathbb{E}\,|\boldsymbol{x}_n|^p < \infty$. Then, the sequence $\{\boldsymbol{x}_n\}$ is said to converge in the $p-$th mean to the random variable $\boldsymbol{x}$ if

(convergence in $p-$th mean)

$$\boldsymbol{x}_n \xrightarrow{L^p} \boldsymbol{x} \iff \lim_{n\to\infty} \mathbb{E}\,|\boldsymbol{x}_n - \boldsymbol{x}|^p = 0 \tag{3.228}$$

The notation L^p above an arrow is used to refer to this notion of convergence. Two special cases are common: $p = 2$ corresponds to mean-square convergence and $p = 1$ corresponds to convergence in the mean. It is easy to verify that convergence in the $p-$th mean implies convergence in probability — see Prob. 3.50.

3.B CONCENTRATION INEQUALITIES

The Markov and Chebyshev bounds are examples of *concentration inequalities*, which help bound the deviation of a random variable (or combinations of random variables) away from certain values (typically their means):

$$\mathbb{P}(\boldsymbol{x} \geq \alpha) \leq \mathbb{E}\,\boldsymbol{x}/\alpha, \quad \boldsymbol{x} \geq 0, \quad \alpha > 0, \qquad (\textbf{Markov inequality}) \tag{3.229a}$$

$$\mathbb{P}(|\boldsymbol{x} - \mathbb{E}\,\boldsymbol{x}| \geq \delta) \;\leq\; \sigma_x^2/\delta^2, \quad \delta > 0, \qquad (\textbf{Chebyshev inequality}) \tag{3.229b}$$

In this appendix, we describe three other famous inequalities known as Azuma inequality, Hoeffding inequality, and McDiarmid inequality, which provide bounds on the probability of the sum of a collection of random variables deviating from their mean.

The Hoeffding and McDiarmid inequalities play an important role in the analysis of learning algorithms and will be used in the derivation of generalization bounds in future Chapter 64. There are of course other concentration inequalities but we will limit our discussion to those that are most relevant to our treatment in this text.

Hoeffding inequality

We first establish Hoeffding inequality from Prob. 3.54 and the supporting Hoeffding lemma, motivated by arguments from Hoeffding (1963), Serfling (1974), Boucheron, Lugosi, and Bousquet (2004), Massart (207), Boucheron, Lugosi, and Massart (2013), Mohri, Rostamizadeh, and Talwalkar (2018), Vershynin (2018), and Wainwright (2019).

Hoeffding inequality (Hoeffding (1963)). *Let* $\{\boldsymbol{x}_n,\ n=1,\ldots,N\}$ *denote* N *independent scalar random variables, with each variable lying within an interval of the form* $a_n \le \boldsymbol{x}_n \le b_n$*, with endpoints denoted by* $\{a_n, b_n\}$*. Let*

$$\boldsymbol{S}_N \triangleq \sum_{n=1}^{N} \boldsymbol{x}_n, \qquad \Delta \triangleq \sum_{n=1}^{N} (b_n - a_n)^2 \tag{3.230}$$

denote the sum of the random variables and the sum of the squared lengths of their respective intervals. The Hoeffding inequality asserts that for any $\delta > 0$*:*

$$\mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \ge \delta\Big) \le e^{-2\delta^2/\Delta} \tag{3.231a}$$

$$\mathbb{P}\Big(\Big|\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N\Big| \ge \delta\Big) \le 2e^{-2\delta^2/\Delta} \tag{3.231b}$$

The above inequalities can also be restated in terms of sample means as opposed to sums. It is straightforward to verify that

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\,\boldsymbol{x}_n \ge \delta\right) \le e^{-2N^2\delta^2/\Delta} \tag{3.232a}$$

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\,\boldsymbol{x}_n\right| \ge \delta\right) \le 2e^{-2N^2\delta^2/\Delta} \tag{3.232b}$$

One of the tools we employ to establish the inequalities is Hoeffding lemma, which is stated next along with one traditional derivation — see, e.g., Boucheron, Lugosi, and Bousquet (2004), Chung and Lu (2006a,b), Alon and Spencer (2008), Boucheron, Lugosi, and Massart (2013), and Mohri, Rostamizadeh, and Talwalkar (2018).

Hoeffding lemma (Hoeffding (1963)). *Consider a scalar random variable* $\boldsymbol{y} \in [a, b]$*. Then for any* t*, we have*

$$\mathbb{E}\,e^{t(\boldsymbol{y}-\mathbb{E}\,\boldsymbol{y})} \le e^{t^2(b-a)^2/8} \tag{3.233}$$

Proof: We start by noting that the exponential function $f(x) = e^x$ is convex and, therefore, it holds that

$$e^{t\boldsymbol{y}} \le \left(\frac{b-\boldsymbol{y}}{b-a}\right)e^{ta} + \left(\frac{\boldsymbol{y}-a}{b-a}\right)e^{tb} \tag{3.234}$$

where the nonnegative coefficients $(b-\boldsymbol{y})/(b-a)$ and $(\boldsymbol{y}-a)/(b-a)$ add up to one. It follows that

$$\begin{aligned}
e^{-t\mathbb{E}\,\boldsymbol{y}}\,\mathbb{E}\,e^{t\boldsymbol{y}} &\leq e^{-t\mathbb{E}\,\boldsymbol{y}} \times \left\{\left(\frac{b-\mathbb{E}\,\boldsymbol{y}}{b-a}\right)e^{ta} + \left(\frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a}\right)e^{tb}\right\} \\
&= e^{-t\mathbb{E}\,\boldsymbol{y}} \times e^{ta} \times \left\{1 - \frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a} + \frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a}e^{t(b-a)}\right\} \\
&= \exp\left\{-t\,(b-a)\,\frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a}\right\} \times \left\{1 - \frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a} + \left(\frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a}\right)e^{t(b-a)}\right\} \\
&= \exp\left\{-t\,(b-a)\,\frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a} + \ln\left(1 - \frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a} + \frac{\mathbb{E}\,\boldsymbol{y}-a}{b-a}e^{t(b-a)}\right)\right\} \\
&\triangleq \exp\left\{-hp + \ln\left(1-p+pe^{h}\right)\right\} \\
&\triangleq e^{L(h)}
\end{aligned} \tag{3.235}$$

where we introduced the quantities:

$$h \triangleq t(b-a) \geq 0, \qquad p \triangleq \frac{\mathbb{E}\boldsymbol{y}-a}{b-a}, \qquad L(h) \triangleq -hp + \ln(1-p+pe^{h}) \tag{3.236}$$

We denote the first and second-order derivatives of $L(h)$ with respect to h by:

$$L'(h) = -p + pe^{h}\frac{1}{1-p+pe^{h}} \tag{3.237a}$$

$$L''(h) = \frac{(1-p)pe^{h}}{(1-p+pe^{h})^2} \tag{3.237b}$$

and note that $L(0) = L'(0) = 0$ and $L''(h) \leq 1/4$. This last inequality follows from the following equivalent statements:

$$\begin{aligned}
\frac{(1-p)pe^{h}}{(1-p+pe^{h})^2} \leq \frac{1}{4} &\iff 4(1-p)pe^{h} \leq (1-p+pe^{h})^2 \\
&\iff 0 \leq (1-p-pe^{h})^2
\end{aligned} \tag{3.238}$$

and the fact that the last statement is obviously true. Now, since $h \geq 0$, we expand $L(h)$ around $h = 0$ and use the mean-value theorem to conclude that there exists a nonnegative value c between 0 and h such that

$$\begin{aligned}
L(h) &= L(0) + L'(0)h + \frac{L''(c)}{2}h^2 \\
&\leq h^2/8 \\
&= t^2(b-a)^2/8
\end{aligned} \tag{3.239}$$

and, consequently,

$$\mathbb{E}\,e^{t(\boldsymbol{y}-\mathbb{E}\,\boldsymbol{y})} \leq e^{L(h)} \leq e^{t^2(b-a)^2/8} \tag{3.240}$$

as claimed. ■

We can now return to establish Hoeffding inequality (3.231b).

Proof of Hoeffding inequality (3.231b): To begin with, we note that for any positive

scalar $s > 0$, it holds that:

$$\begin{aligned}\mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \geq \delta\Big) &= \mathbb{P}\left(e^{s(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N)} \geq e^{s\delta}\right) \\ &\leq e^{-s\delta}\,\mathbb{E}\left(e^{s(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N)}\right) \end{aligned} \tag{3.241}$$

where we used Markov inequality from Prob. 3.17, namely, the fact that for any non-negative real-valued random variable $\boldsymbol{x}$, it holds that $\mathbb{P}(\boldsymbol{x} \geq \alpha) \leq \mathbb{E}\,\boldsymbol{x}/\alpha$. Now, from the definition of $\boldsymbol{S}_N$ in (3.230) and the independence of the $\{\boldsymbol{x}(n)\}$ we get:

$$\begin{aligned} e^{-s\delta}\,\mathbb{E}\left(e^{s(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N)}\right) &= e^{-s\delta}\,\mathbb{E}\left(e^{s\left(\sum_{n=1}^N \boldsymbol{x}_n - \mathbb{E}\,\boldsymbol{x}_n\right)}\right) \\ &= e^{-s\delta}\,\mathbb{E}\left(\prod_{n=1}^N e^{s(\boldsymbol{x}_n - \mathbb{E}\,\boldsymbol{x}_n)}\right) \\ &= e^{-s\delta}\prod_{n=1}^N \mathbb{E}\left(e^{s(\boldsymbol{x}_n - \mathbb{E}\,\boldsymbol{x}_n)}\right) \end{aligned} \tag{3.242}$$

so that

$$\mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \geq \delta\Big) \leq e^{-s\delta}\prod_{n=1}^N \mathbb{E}\left(e^{s(\boldsymbol{x}_n - \mathbb{E}\,\boldsymbol{x}_n)}\right) \tag{3.243}$$

To continue, we call upon result (3.233) from Hoeffding lemma to conclude that

$$\begin{aligned}\mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \geq \delta\Big) &\leq e^{-s\delta}\prod_{n=1}^N e^{\frac{s^2}{8}(b_n - a_n)^2} \\ &= e^{-s\delta}\,e^{\frac{s^2}{8}\sum_{n=1}^N (b_n - a_n)^2} \\ &= e^{-s\delta}\,e^{\frac{s^2\Delta}{8}} \end{aligned} \tag{3.244}$$

We can tighten the upper bound by selecting the value of s that minimizes the exponent, $-s\delta + s^2\Delta/8$, which is given by $s = 4\delta/\Delta$. Therefore, we get

$$\mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \geq \delta\Big) \leq e^{-2\delta^2/\Delta} \tag{3.245}$$

Following similar arguments, we can also get

$$\mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \leq -\delta\Big) \;=\; \mathbb{P}\Big(-[\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N] \geq \delta\Big) \;\leq\; e^{-2\delta^2/\Delta} \tag{3.246}$$

We then arrive at

$$\begin{aligned}&\mathbb{P}\Big(|\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N| \geq \delta\Big) \\ &= \mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \leq -\delta\Big) + \mathbb{P}\Big(\boldsymbol{S}_N - \mathbb{E}\,\boldsymbol{S}_N \geq \delta\Big) \\ &\leq 2e^{-2\delta^2/\Delta} \end{aligned} \tag{3.247}$$

■

Azuma and McDiarmid Inequalities

The Hoeffding inequalities (3.232a)–(3.232b) provide bounds for the deviation of the sample average function:

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \;\triangleq\; \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n \tag{3.248}$$

away from its mean. The results can be extended to other functions with "*bounded variation*." These are again functions of the $N-$variables $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, except that if any of the variables is changed, say, from $\boldsymbol{x}_m$ to $\boldsymbol{x}'_m$, then the variation in the function remains bounded:

(**function with bounded variations**)

$$\sup_{x_1,\ldots,x_N,x'_m} \left\{\left|f(\boldsymbol{x}_{n\neq m}, \boldsymbol{x}_m) - f(\boldsymbol{x}_{n\neq m}, \boldsymbol{x}'_m)\right|\right\} \leq c_m, \quad \forall\, m = 1, 2, \ldots, N \tag{3.249}$$

For such functions, the Hoeffding inequalities extend to the *McDiarmid inequalities* stated in (3.259a)–(3.259b) further ahead. These results were proven by McDiarmid (1989); see also Ledoux (2001), Chung and Lu (2006a,b), Alon and Spencer (2008), Boucheron, Lugosi, and Massart (2013), Mohri, Rostamizadeh, and Talwalkar (2018), and Wainwright (2019). Motivated by the presentation in these references, we provide one classical derivation that relies on the use of the Azuma inequality, which we motivate first.

Consider two sequences of random variables $\{\boldsymbol{y}_n, \boldsymbol{x}_n\}$ for $n \geq 1$, where each $\boldsymbol{y}_n$ is a function of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. We say that the sequence $\{\boldsymbol{y}_n\}$ is a *martingale difference* relative to the sequence $\{\boldsymbol{x}_n\}$ if the following property holds for every n:

$$\mathbb{E}\left(\boldsymbol{y}_n \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}\right) = 0, \quad (\textbf{martingale difference}) \tag{3.250}$$

Azuma inequality (Azuma (1967)). *Let $\{\boldsymbol{y}_n,\ n \geq 1\}$ be a martingale difference relative to another sequence $\{\boldsymbol{x}_n,\ n \geq 1\}$, and assume there exist random variables $\{\boldsymbol{z}_n\}$ and constants $\{c_n\}$ such that $\boldsymbol{z}_n \leq \boldsymbol{y}_n \leq \boldsymbol{z}_n + c$ for all n. Let $\Delta' = \sum_{n=1}^N c_n^2$. The Azuma inequality asserts that for any $\delta > 0$:*

$$\mathbb{P}\Big(\sum_{n=1}^N \boldsymbol{y}_n \geq \delta\Big) \leq e^{-2\delta^2/\Delta'} \tag{3.251a}$$

$$\mathbb{P}\Big(\sum_{n=1}^N \boldsymbol{y}_n \leq -\delta\Big) \leq e^{-2\delta^2/\Delta'} \tag{3.251b}$$

Proof: It is sufficient to establish one of the inequalities. We follow an argument similar to Chung and Lu (2006b) and Mohri, Rostamizadeh, and Talwalkar (2018). Introduce the random variable $\boldsymbol{S}_N = \sum_{n=1}^N \boldsymbol{y}_n$, which satisfies $\boldsymbol{S}_N = \boldsymbol{S}_{N-1} + \boldsymbol{y}_N$. For any $s > 0$ we have

$$\begin{aligned}
\mathbb{P}(\boldsymbol{S}_N \geq \delta) &= \mathbb{P}(e^{s\boldsymbol{S}_N} \geq e^{s\delta}) \\
&\leq \frac{\mathbb{E}\, e^{s\boldsymbol{S}_N}}{e^{s\delta}}, \quad \text{(using Markov inequality from Prob. 3.17)} \\
&= e^{-s\delta} \times \mathbb{E}\, e^{s(\boldsymbol{S}_{N-1} + \boldsymbol{y}_N)} \\
&= e^{-s\delta} \times \mathbb{E}\left\{\mathbb{E}\left(e^{s(\boldsymbol{S}_{N-1}+\boldsymbol{y}_N)} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}\right)\right\} \\
&\overset{(a)}{=} e^{-s\delta} \times \mathbb{E}\left\{e^{s\boldsymbol{S}_{N-1}}\, \mathbb{E}\left(e^{s\boldsymbol{y}_N} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}\right)\right\} \\
&\overset{(b)}{\leq} e^{-s\delta} \times \mathbb{E}\, e^{s\boldsymbol{S}_{N-1}} \times e^{s^2 c_N^2/8}
\end{aligned} \tag{3.252}$$

where step (a) is because $\boldsymbol{S}_{N-1}$ is solely a function of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N-1}\}$, and step (b) uses the result of Prob. 3.57. We therefore arrive at the inequality recursion:

$$\mathbb{E}\, e^{s\boldsymbol{S}_N} \leq \mathbb{E}\, e^{s\boldsymbol{S}_{N-1}} \times e^{s^2 c_N^2/8} \tag{3.253}$$

Iterating starting from $\boldsymbol{S}_0 = 0$ we get

$$\mathbb{E}\, e^{s\boldsymbol{S}_N} \leq e^{\sum_{n=1}^{N} s^2 c_n^2} \tag{3.254}$$

and, hence,

$$\mathbb{P}(\boldsymbol{S}_N \geq \delta) \leq e^{-s\delta} \times e^{s^2 \sum_{n=1}^{N} c_n^2/8} \tag{3.255}$$

We can minimize the upper bound over s and select

$$s = 4\delta \Big/ \sum_{n=1}^{N} c_n^2/8 \tag{3.256}$$

Substituting into the right-hand side of (3.255) gives

$$\mathbb{P}(\boldsymbol{S}_N \geq \delta) \leq e^{-2\delta^2/\sum_{n=1}^{N} c_n^2} \tag{3.257}$$

and the desired result follows.

■

We are now ready to state the McDiarmid inequality.

McDiarmid inequality (McDiarmid (1989)). *Let $\{\boldsymbol{x}_n,\ n = 1, \ldots, N\}$ denote N independent scalar random variables, and let $f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$ be any function with bounded variation as in (3.249). Let*

$$\Delta' \triangleq \sum_{m=1}^{N} c_m^2 \tag{3.258}$$

The McDiarmid inequality asserts that for any $\delta > 0$:

$$\mathbb{P}\Big(f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) - \mathbb{E}\, f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \geq \delta \Big) \leq e^{-2\delta^2/\Delta'} \tag{3.259a}$$

$$\mathbb{P}\Big(|f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) - \mathbb{E}\, f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)| \geq \delta \Big) \leq 2e^{-2\delta^2/\Delta'} \tag{3.259b}$$

Proof: It is sufficient to establish one of the inequalities. Introduce the following random variables, where the expression for $\boldsymbol{y}_n$ is written in two equivalent forms:

$$\boldsymbol{S} \triangleq f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) - \mathbb{E}\, f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \tag{3.260a}$$

$$\begin{aligned} \boldsymbol{y}_n &\triangleq \mathbb{E}\left(\boldsymbol{S} \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\right) - \mathbb{E}\left(\boldsymbol{S} \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}\right), \quad n \geq 1 \\ &= \mathbb{E}\left(f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\right) - \mathbb{E}\left(f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}\right) \end{aligned} \tag{3.260b}$$

It is clear that $\boldsymbol{S} = \sum_{n=1}^{N} \boldsymbol{y}_n$ and $\mathbb{E}\,(\boldsymbol{y}_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}) = 0$. The latter result shows that the sequence $\{\boldsymbol{y}_n\}$ defines a martingale difference relative to the sequence $\{\boldsymbol{x}_n\}$. Moreover, the bounded variation property on the function $f(\cdot)$ translates into bounds on each $\boldsymbol{y}_n$ as follows. Let

$$a_n \triangleq \inf_{x} \left\{ \mathbb{E}\left(f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, x\right) - \mathbb{E}\left(f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\right) \right\} \tag{3.261}$$

Then, each $\boldsymbol{y}_n$ satisfies $a_n \leq \boldsymbol{y}_n \leq a_n + c_n$. We can now apply Azuma inequality (3.251a) to get (3.259a).

■

REFERENCES

Alon, N. and J. H. Spencer (2008), *The Probabilistic Method*, 3rd edition, Wiley Series in Discrete Mathematics and Optimization, NY.

Ash, R. B. (2008), *Basic Probability Theory* , Dover Publications, NY.

Azuma, K. (1967), "Weighted sums of certain dependent random variables," *Tôhoku Mathematical Journal*, vol. 19, no. 3, pp. 357–367.

Bass, R. F. (2011), *Stochastic Processes*, Cambridge University Press.

Bayes, T. and R. Price (1763), "An essay towards solving a problem in the doctrine of chances," Bayes's article communicated by R. Price and published posthumously in the *Philosophical Trans. Royal Society of London*, vol. 53, pp. 370–418.

Berg, H. C. (1993), *Random Walks in Biology*, expanded edition, Princeton University Press, NJ.

Bernoulli, J. (1713), *Ars Conjectandi*, Chapter 4, Thurneysen Brothers, Basel. Book published eight years after the author's death. Enlgish translation by E. Sylla available as *The Art of Conjecturing*, Johns Hopkins University Press.

Bernshtein, S. N. (1945), "Chebyshev's work on the theory of probability," *Akademiya Nauk SSSR*, pp. 43–68, Moscow–Leningrad.

Bienaymé, I. J. (1853), "Considérations al'appui de la découverte de Laplace," *Comptes Rendus de l'Académie des Sciences*, vol. 37, pp. 309–324.

Billingsley, P. (1986), *Probability and Measure*, 2nd edition, Wiley, NY.

Billingsley, P. (1999), *Convergence of Probability Measures*, 2nd edition, Wiley.

Bochner, S. (1955), *Harmonic Analysis and the Theory of Probability*, University of California Press.

Boltzmann, L. (1877), "Uber die beziehung dem zweiten haubtsatze der mechanischen warmetheorie und der wahrscheinlichkeitsrechnung respektive den Satzen uber das warmegleichgewicht," *Wiener Berichte*, vol. 76, pp. 373–435. in WA II, paper 42.

Boltzmann, L. (1909), *Wissenschaftliche Abhandlungen*, vols. I, II, and III, F. Hasenöhrl, *Ed.*, Barth, Leipzig. Reissued by Chelsea, NY, 1969.

Borel, E. (1909), "Les probabilités dénombrables et leurs applications arithmétique," *Rend. Circ. Mat. Palermo*, vol. 2, no. 27, pp. 247–271.

Boucheron, S., G. Lugosi, and O. Bousquet (2004), "Concentration Inequalities," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch, *Eds.*, pp. 208–240, Springer.

Boucheron, S., G. Lugosi, and P. Massart (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

Brown, R. (1828),"A brief account of microscopical observations made in the months of June, July, and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies," *Edinburgh New Philos. J.*, vol. 5, pp. 358–371; reprinted in *Philos. Mag.*, vol. 4, pp. 161–173. Addendum, "Additional remarks on active molecules," appears in *Edinburgh J. Sci*, vol. 1, p. 314, 1829.

Brown, R. (1866), *The Miscellaneous Botanical Works of Robert Brown*, vols. 1 and 2, Robert Hardwicke, London.

Chebyshev, P. (1867), "Des valeurs moyennes," *J. de Mathématiques Pures et Appliquées*, vol. 2, no. 12, pp. 177–184.

Chernoff, H. (1952), "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations, " *Annals of Mathematical Statistics*, vol. 23, pp. 493–507.

Chung, K. L. (2000), *A Course in Probability Theory*, 2nd edition, Academic Press.

Chung, F. and L. Lu (2006a), *Complex Graphs and Networks*, vol. 107, Regional Conference Series in Mathematics, American Mathematical Society (AMS).
Chung, F. and L. Lu (2006b), "Concentration inequalities and martingale inequalities: A survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127.
Dale, A. I. (2003), *Most Honourable Remembrance: The Life and Work of Thomas Bayes*, Springer-Verlag, NY.
Davidson, J. (1994), *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.
Doob, J. L (1996), "The development of rigor in mathematical probability (1900-1950)," *Amer. Math. Monthly*, vol. 103, pp. 586–595.
Drineas, P., R. Kannan, and M. W. Mahoney (2006a), "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," *SIAM Journal on Computing*, vol. 36, pp. 132–157.
Dubois, J.-M., O. Gilles, and B. Rouzaire-Dubois (2009), "The Boltzmann equation in molecular biology," *Progress in Biophysics and Molecular Biology*, vol. 99, pp. 87–93.
Dudley, R. M. (2002), *Real Analysis and Probability*, Cambridge University Press.
Durrett, R. (1996), *Probability: Theory and Examples*, 2nd edition, Duxbury Press, London. Fifth edition published by Cambridge University Press, 2019.
Edwards, A. W. F. (1986), "Is the reference in Hartley (1749) to Bayesian inference?" *The American Statistician*, vol. 40, no. 2, pp.109–110.
Einstein, A. (1905), "Uber die von der molekularkinetischen theorie der warme geforderte bewegung von in ruhenden flussigkeiten suspendierten teilchenm" *Annalen der Physik*, vol. 322, no. 8, pp. 549–560.
Feinberg, S. E. (2003), "When did Bayesian inference become Bayesian?" *Bayesian Analysis*, vol. 1, no. 1, pp. 1–37.
Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd edition, Wiley, NY.
Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, vol. 2, 3rd edition, Wiley, NY.
Fischer, H. (2011), *A History of the Central Limit Theorem*, Springer, NY.
Gallager, R. G. (2014), *Stochastic Processes: Theory for Applications*, Cambridge University Press.
Gibbs, J. W. (1902), *Elementary Principles in Statistical Mechanics*, Charles Scribner's Sons, NY.
Gnedenko, B. V. (1998), *Theory of Probability*, 6th edition, CRC.
Grimmett, G. and D. Stirzaker (2001), *Probability and Random Processes*, 3rd edition, Oxford University Press.
Hald, A. (1998), *A History of Mathematical Statistics From 1750 to 1930*, Wiley, NY.
Hardy, G. H., J. E. Littlewood, and G. Pólya (1934), *Inequalities*, Cambridge University Press.
Hill, T. L. (1987), *Statistical Mechanics: Principles and Selected Applications*, Dover, NY.
Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *Journal Amer. Stat. Assoc.*, vol. 58, pp. 13–30.
Huang, K. (2005), *Lectures on Statistical Physics and Protein Folding*, World Scientific.
Ingenhousz, J. (1784), "Remarks on the use of the magnifying glass," *Journal de Physique*, vol. II, pp. 122–126. An English translation appears in P. W. van der Pas,"The discovery of the Brownian motion" *Scientiarum Historia*, vol. 13, pp. 27–35.
Katznelson, Y. (2004), *An Introduction to Harmonic Analysis,* 3rd edition, Cambridge University Press.
Knuth, D. (1997), *The Art of Computer Programming*, 3rd edition, Addison-Wesley.
Kolmogorov, A. N. (1927) "Sur la loi forte des grands nombres," *C. R. Acad. Sci.*, Ser. I Math, no. 191, pp. 910–912, Paris.
Kolmogorov, A. N. (1931), "Ueber die analytischen methoden der wahrscheinlichkeitsrechnung," *Math. Ann.*, vol. 104, pp. 415–458.
Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, vol. 2, no.

3, Springer, Berlin. An English translation by N. Morrison exists and is entitled *Foundations of the Theory of Probability*, 2nd edition, Chelsea, NY, 1956.

Kolmogorov, A. N. (1960), *Foundations of the Theory of Probability*, 2nd edition, Chelsea Pub Co.

Landau, L. D. and E. M. Lifshitz (1980), *Statistical Physics*, 3rd edition, Butterworth-Heinemann.

Laplace, P. S. (1774), "Mémoire sur la probabilité des causes par les événements," *Mém. Acad. R. Sci. de MI (Savants étrangers)*, vol. 4, pp. 621–656. See also *Oeuvres Complètes de Laplace*, vol. 8, pp. 27–65 published by the L'Académie des Sciences, Paris, during the period 1878–1912. Translated by S. M. Sitgler, *Statistical Science*, vol. 1, no. 3, pp. 366–367.

Lawler, G. F. and V. Limic (2010), *Random Walk: A Modern Introduction*, Cambridge University Press.

Ledoux, M. (2001), *The Concentration of Measure Phenomenon*, vol. 89, Mathematical Surveys and Monographs, American Mathematical Society (AMS).

Leon-Garcia, A. (2008), *Probability, Statistics, and Random Processes For Electrical Engineering,* 3rd edition, Prentice Hall, NJ.

Lukacs, E. (1970), *Characteristic Functions*, 2nd edition, Charles Griffin & Co.

Markov, A. A. (1884), *On Certain Applications of Algebraic Continued Fractions*, Ph.D. dissertation, St. Petersburg, Russia.

Massart, P. (2007), *Concentration Inequalities and Model Selection*, Springer, NY.

McDiarmid, C. (1989), "On the method of bounded differences," pp. 148–188, in *Surveys in Combinatorics*, J. Siemons, *Ed.*, Cambridge University Press.

Morters, P. and Y. Peres (2010), *Brownian Motion*, Cambridge University Press.

Okamoto, M. (1958), "Some inequalities relating to the partial sum of binomial probabilities," *Annals Inst. Stat. Math.*, vol. 10, pp. 29–35.

Ondar, K. O. (1981), *The Correspondence Between A. A. Markov and A. A. Chuprov on the Theory of Probability and Mathematical Statistics*, Springer, NY.

Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes (1997), "Theory of protein folding: The energy landscape perspective," *Annu. Rev. Phys. Chem.*, vol. 48, pp. 545–600.

Oppenheim, A. V., R. W. Schafer, and J. R. Buck (2009), *Discrete-Time Signal Processing*, 3rd edition, Prentice Hall, NJ.

Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes,* 3rd edition, McGraw-Hill, NY.

Pathria, R. K. and P. D. Beale (2011), *Statistical Mechanics*, 3rd edition, Academic Press.

Pearl, J. (1995), "Causal diagrams for empirical research," *Biometrika*, vol. 82, pp. 669–710.

Pearl, J. *2000*, *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Pearle, P., B. Collett, K. Bart, D. Bilderback, D. Newman, and S. Samuels (2010), "What Brown saw and you can too," *Amer. J. Phys.*, vol.78, pp. 1278–1289.

Picinbono, B. (1993), *Random Signals and Systems*, Prentice Hall, NJ.

Phillips, R., J. Kondev, J. Theriot, H. Garcia, and J. Kondev (2012), *Physical Biology of the Cell*, 2nd edition, Garland Science.

Poisson, S. D. (1837), *Probabilité des Jugements en Matière Criminelle et en Matière Civile*, Bachelier, Paris.

Prokhorov, Y. V. (2011), "Strong law of large numbers," Encyclopedia of Mathematics. Available online at http://encyclopediaofmath.org

Rogers, L. C. G. and D. Williams (2000), *Diffusions, Markov Processes, and Martingales*, Cambridge University Press.

Sayed, A. H. (2003), *Fundamentals of Adaptive Filtering*, Wiley, NJ.

Sayed, A. H. (2008), *Adaptive Filters*, Wiley, NJ.

Santana, R., P. Larranaga, and J. A. Lozano (2008), "Protein folding in simplified

models with estimation of distribution algorithms," *IEEE Trans. Evolut. Comput.*, vol. 12, no. 4, pp. 418–438.
Seneta, E. (2013), "A tricentenary history of the law of large numbers," *Bernoulli*, vol. 19, no. 4, pp. 1088–1121.
Serfling, R. J. (1974), "Probability inequalities for the sum in sampling without replacement," *The Annals of Statistics*, vol. 2, no. 1, pp. 39–48.
Shafer, G. and V. Vovk (2006), "The sources of Kolmogorov's Grundbegriffe," *Statistical Science*, vol. 21, no. 1, pp. 70–98.
Shiryayev, A. N. (1984), *Probability*, Springer, NY.
Slutsky, E. (1925), "Über stochastische Asymptoten und Grenzwerte," *Metron*, in German, vol. 5, no. 3, pp. 3–89.
Smoluchowski, M. (1906), "Zur kinetischen theorie der Brownschen molekularbewegung und der suspensionen," *Annalen der Physik*, vol. 326, no. 14, pp. 756–780.
Stark, H. and J. W. Woods (1994), *Probability, Random Processes, and Estimation Theory for Engineers*, 2nd edition, Prentice Hall, NJ.
Stein, E. M. and R. Shakarchi (2003), *Fourier Analysis: An Introduction*, Princeton University Press.
Stigler, S. M. (1983), "Who discovered Bayes' theorem?" *The American Statistician*, vol. 37, no. 4, pp. 290–296.
Sutherland, W. (1905), "A dynamical theory of diffusion for non-electrolytes and the molecular mass of albumin," *Phil. Mag.*, vol. 9, pp. 781–785.
Tankard, J. W. (1984), *The Statistical Pioneers*, Schenkman Books.
Tolman, R. C. (2010), *The Principles of Statistical Mechanics*, Dover Books.
Uffink, J. (2014), "Boltzmann's work in statistical physics," *The Stanford Encyclopedia of Philosophy*, Fall 2014 edition,, E. N. Zalta, *Ed.*, available online at the location http://plato.stanford.edu/archives/fall2014/entries/statphys-Boltzmann/
van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge University Press.
van der Pas, P. W. (1971), "The discovery of the Brownian motion" *Scientiarum Historia*, vol. 13, p. 27–35.
Vershynin, R. (2018), *High-Dimensional Probability*, Cambridge University Press.
Vetterli, M., J. Kovacevic, and V. K. Goyal (2014), *Foundations of Signal Processing*, Cambridge University Press.
Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press.
Weiss, N. A. (2005), *A Course in Probability*, Addison-Wesley, MA.