Probability and Random Processes for Electrical Engineers

John A. Gubner
University of Wisconsin–Madison

Discrete Random Variables

Bernoulli(p)

$\wp(X=1)=p, \quad \wp(X=0)=1-p.$
$\mathsf{E}[X]=p, \quad \mathsf{var}(X)=p(1-p), \quad G_X(z)=(1-p)+pz.$

---

binomial(n,p)

$$\wp(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad k=0,\ldots,n.$$

$\mathsf{E}[X]=np, \quad \mathsf{var}(X)=np(1-p), \quad G_X(z)=[(1-p)+pz]^n.$

---

geometric$_0$(p)

$\wp(X=k)=(1-p)p^k, \quad k=0,1,2,\ldots.$

$\mathsf{E}[X]=\dfrac{p}{1-p}, \quad \mathsf{var}(X)=\dfrac{p}{(1-p)^2}, \quad G_X(z)=\dfrac{1-p}{1-pz}.$

---

geometric$_1$(p)

$\wp(X=k)=(1-p)p^{k-1}, \quad k=1,2,3,\ldots.$

$\mathsf{E}[X]=\dfrac{1}{1-p}, \quad \mathsf{var}(X)=\dfrac{p}{(1-p)^2}, \quad G_X(z)=\dfrac{(1-p)z}{1-pz}.$

---

negative binomial or Pascal(m,p)

$$\wp(X=k)=\binom{k-1}{m-1}(1-p)^m p^{k-m}, \ k=m,m+1,\ldots.$$

$$\mathsf{E}[X]=\frac{m}{1-p}, \quad \mathsf{var}(X)=\frac{mp}{(1-p)^2}, \quad G_X(z)=\left[\frac{(1-p)z}{1-pz}\right]^m.$$

Note that Pascal$(1,p)$ is the same as geometric$_1(p)$.

Poisson(λ)

$$\wp(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, \quad k=0,1,\ldots.$$

$\mathsf{E}[X]=\lambda, \quad \mathsf{var}(X)=\lambda, \quad G_X(z)=e^{\lambda(z-1)}.$

Fourier Transforms

Fourier Transform

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{-j2\pi ft}\, dt$$

Inversion Formula

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{j2\pi ft}\, df$$

$h(t)$	$H(f)$
$I_{[-T,T]}(t)$	$2T\dfrac{\sin(2\pi Tf)}{2\pi Tf}$
$2W\dfrac{\sin(2\pi Wt)}{2\pi Wt}$	$I_{[-W,W]}(f)$
$(1-\lvert t\rvert/T)\, I_{[-T,T]}(t)$	$T\left[\dfrac{\sin(\pi Tf)}{\pi Tf}\right]^2$
$W\left[\dfrac{\sin(\pi Wt)}{\pi Wt}\right]^2$	$(1-\lvert f\rvert/W) I_{[-W,W]}(f)$
$e^{-\lambda t}u(t)$	$\dfrac{1}{\lambda + j2\pi f}$
$e^{-\lambda\lvert t\rvert}$	$\dfrac{2\lambda}{\lambda^2 + (2\pi f)^2}$
$\dfrac{\lambda}{\lambda^2 + t^2}$	$\pi\, e^{-2\pi\lambda\lvert f\rvert}$
$e^{-(t/\sigma)^2/2}$	$\sqrt{2\pi}\,\sigma\, e^{-\sigma^2(2\pi f)^2/2}$

Preface

Intended Audience

This book contains enough material to serve as a text for a two-course sequence in probability and random processes for electrical engineers. It is also useful as a reference by practicing engineers.

For students with no background in probability and random processes, a first course can be offered either at the undergraduate level to talented juniors and seniors, or at the graduate level. The prerequisite is the usual undergraduate electrical engineering course on signals and systems, e.g., Haykin and Van Veen [18] or Oppenheim and Willsky [30] (see the Bibliography at the end of the book).

A second course can be offered at the graduate level. The additional prerequisite is some familiarity with linear algebra; e.g., matrix-vector multiplication, determinants, and matrix inverses. Because of the special attention paid to complex-valued Gaussian random vectors and related random variables, the text will be of particular interest to students in wireless communications. Additionally, the last chapter, which focuses on self-similar processes, long-range dependence, and aggregation, will be very useful for students in communication networks who are interested modeling Internet traffic.

Material for a First Course

In a first course, Chapters 1–5 would make up the core of any offering. These chapters cover the basics of probability and discrete and continuous random variables. Following Chapter 5, additional topics such as wide-sense stationary processes (Sections 6.1–6.5), the Poisson process (Section 8.1), discrete-time Markov chains (Section 9.1), and confidence intervals (Sections 12.1–12.4) can also be included. These topics can be covered independently of each other, in any order, except that Problem 15 in Chapter 12 refers to the Poisson process.

Material for a Second Course

In a second course, Chapters 7–8 and 10–11 would make up the core, with additional material from Chapters 6, 9, 12, and 13 depending on student preparation and course objectives. For review purposes, it may be helpful at the beginning of the course to assign the more advanced problems from Chapters 1–5 that are marked with a *.

Features

Those parts of the book mentioned above as being suitable for a first course are written at a level appropriate for undergraduates. More advanced problems and sections in these parts of the book are indicated by a *. Those parts of the book not indicated as suitable for a first course are written at a level suitable

for graduate students. Throughout the text, there are numerical superscripts that refer to notes at the end of each chapter. These notes are usually rather technical and address subtleties of the theory.

The last section of each chapter is devoted to problems. The problems are separated by section so that all problems relating to a particular section are clearly indicated. This enables the student to refer to the appropriate part of the text for background relating to particular problems, and it enables the instructor to make up assignments more quickly.

Tables of discrete random variables and of Fourier transform pairs are found inside the front cover. A table of continuous random variables is found inside the back cover.

When cdfs or other functions are encountered that do not have a closed form, MATLAB commands are given for computing them. For a list, see "Matlab" in the index.

The index was compiled as the book was being written. Hence, there are many pointers to specific information. For example, see "noncentral chi-squared random variable."

With the importance of wireless communications, it is vital for students to be comfortable with complex random variables, especially complex Gaussian random variables. Hence, Section 7.5 is devoted to this topic, with special attention to the circularly symmetric complex Gaussian. Also important in this regard are the central and noncentral chi-squared random variables and their square roots, the Rayleigh and Rice random variables. These random variables, as well as related ones such as the beta, F, Nakagami, and student's t, appear in numerous problems in Chapters 3–5 and 7.

Chapter 10 gives an extensive treatment of convergence in mean of order p. Special attention is given to mean-square convergence and the Hilbert space of square-integrable random variables. This allows us to prove the projection theorem, which is important for establishing the existence of certain estimators, and conditional expectation in particular. The Hilbert space setting allows us to define the Wiener integral, which is used in Chapter 13 on advanced topics to construct fractional Brownian motion.

We give an extensive treatment of confidence intervals in Chapter 12, not only for normal data, but also for more arbitrary data via the central limit theorem. The latter is important for any situation in which the data is not normal; e.g., sampling for quality control, election polls, etc. Much of this material can be covered in a first course. However, the last subsection on derivations is more advanced, and uses results from Chapter 7 on Gaussian random vectors.

With the increasing use of self-similar and long-range dependent processes for modeling Internet traffic, we provide in Chapter 13 an introduction to these developments. In particular, fractional Brownian motion is a standard example of a continuous-time self-similar process with stationary increments. Fractional autoregressive integrated moving average (FARIMA) processes are developed as examples of important discrete-time processes in this context.

Table of Contents

CHAPTER 1
Introduction to Probability

If we toss a fair coin many times, then we expect that the fraction of heads should be close to 1/2, and we say that 1/2 is the "probability of heads." In fact, we might try to define the probability of heads to be the limiting value of the fraction of heads as the total number of tosses tends to infinity. The difficulty with this approach is that there is no simple way to guarantee that the desired limit exists.

An alternative approach was developed by **A. N. Kolmogorov** in 1933. Kolmogorov's idea was to start with an axiomatic definition of probability and then deduce results logically as consequences of the axioms. One of the successes of Kolmogorov's approach is that under suitable assumptions, it can be proved that the fraction of heads in a sequence of tosses of a fair coin converges to 1/2 as the number of tosses increases. You will meet a simple version of this result in the **weak law of large numbers** in Chapter 2. Generally speaking, a law of large numbers is a theorem that gives conditions under which the numerical average of a large number of measurements converges in some sense to a probability or other parameter specified by the underlying mathematical model. Other laws of large numbers are discussed in Chapters 10 and 11.

Another celebrated limit result of probability theory is the **central limit theorem**, which you will meet in Chapter 4. The central limit theorem says that when you add up a large number of random perturbations, the overall effect has a **Gaussian** or **normal** distribution. For example, the central limit theorem explains why thermal noise in amplifiers has a Gaussian distribution. The central limit theorem also accounts for the fact that the speed of a particle in an ideal gas has the **Maxwell** distribution.

In addition to the more famous results mentioned above, in this book you will learn the "tools of the trade" of probability and random processes. These include probability mass functions and densities, expectation, transform methods, random processes, filtering of processes by linear time-invariant systems, and more.

Since probability theory relies heavily on the use of set notation and set theory, a brief review of these topics is given in Section 1.1. In Section 1.2, we consider a number of simple physical experiments, and we construct mathematical probability models for them. These models are used to solve several sample problems. Motivated by our specific probability models, in Section 1.3, we introduce the general axioms of probability and several of their consequences. The concepts of statistical independence and conditional probability are introduced in Sections 1.4 and 1.5, respectively. Section 1.6 contains the notes that are referenced in the text by numerical superscripts. These notes are usually rather technical and can be skipped by the beginning student. However, the

notes provide a more in-depth discussion of certain topics that may be of interest to more advanced readers. The chapter concludes with problems for the reader to solve. Problems and sections marked by a * are intended for more advanced readers.

1.1. Review of Set Notation

Let Ω be a set of points. If ω is a point in Ω, we write $\omega \in \Omega$. Let A and B be two collections of points in Ω. If every point in A also belongs to B, we say that A is a **subset** of B, and we denote this by writing $A \subset B$. If $A \subset B$ and $B \subset A$, then we write $A = B$; i.e., two sets are equal if they contain exactly the same points.

Set relationships can be represented graphically in **Venn diagrams**. In these pictures, the whole space Ω is represented by a rectangular region, and subsets of Ω are represented by disks or oval-shaped regions. For example, in Figure 1.1(a), the disk A is completely contained in the oval-shaped region B, thus depicting the relation $A \subset B$.

If $A \subset \Omega$, and $\omega \in \Omega$ does not belong to A, we write $\omega \notin A$. The set of all such ω is called the **complement** of A in Ω; i.e.,

$$A^c := \{\omega \in \Omega : \omega \notin A\}.$$

This is illustrated in Figure 1.1(b), in which the shaded region is the complement of the disk A.

The **empty set** or **null set** of Ω is denoted by $\emptyset$; it contains no points of Ω. Note that for any $A \subset \Omega$, $\emptyset \subset A$. Also, $\Omega^c = \emptyset$.

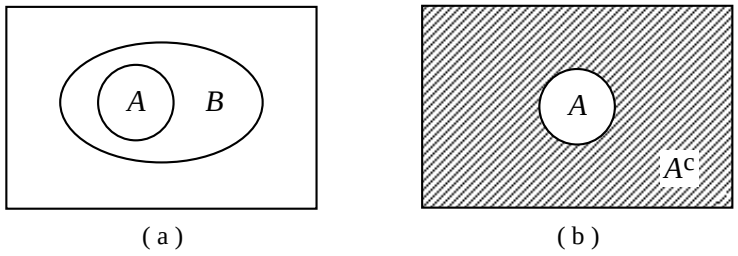

Figure 1.1. (a) Venn diagram of $A \subset B$. (b) The complement of the disk A, denoted by A^c, is the shaded part of the diagram.

The **union** of two subsets A and B is

$$A \cup B := \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$

Here "or" is inclusive; i.e., if $\omega \in A \cup B$, we permit ω to belong to either A or B or both. This is illustrated in Figure 1.2(a), in which the shaded region is the union of the disk A and the oval-shaped region B.

The **intersection** of two subsets A and B is

$$A \cap B := \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\};$$

hence, $\omega \in A \cap B$ if and only if ω belongs to both A and B. This is illustrated in Figure 1.2(b), in which the shaded area is the intersection of the disk A and the oval-shaped region B. The reader should also note the following special case. If $A \subset B$ (recall Figure 1.1(a)), then $A \cap B = A$. In particular, we always have $A \cap \Omega = A$.

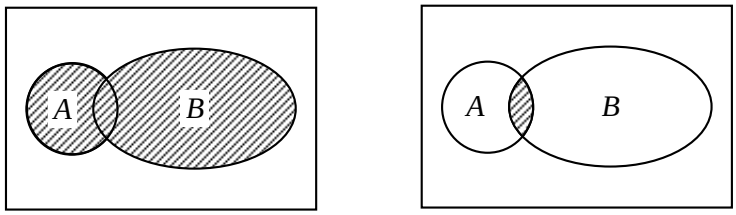

Figure 1.2. (a) The shaded region is $A \cup B$. (b) The shaded region is $A \cap B$.

The **set difference** operation is defined by

$$B \setminus A \;:=\; B \cap A^c,$$

i.e., $B \setminus A$ is the set of $\omega \in B$ that do not belong to A. In Figure 1.3(a), $B \setminus A$ is the shaded part of the oval-shaped region B.

Two subsets A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$; i.e., there is no point in Ω that belongs to both A and B. This condition is depicted in Figure 1.3(b).

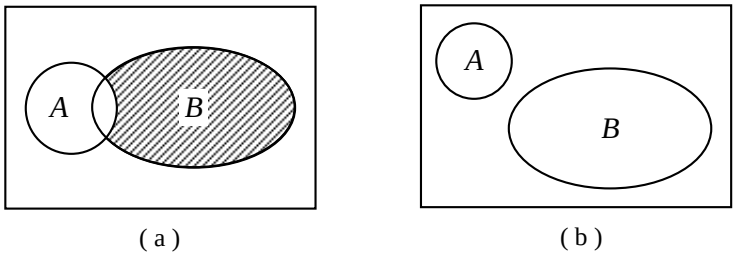

Figure 1.3. (a) The shaded region is $B \setminus A$. (b) Venn diagram of disjoint sets A and B.

Using the preceding definitions, it is easy to see that the following properties hold for subsets A, B, and C of Ω.

The **commutative laws** are

$$A \cup B \;=\; B \cup A \quad \text{and} \quad A \cap B \;=\; B \cap A.$$

The **associative laws** are

$$A \cap (B \cap C) = (A \cap B) \cap C \quad \text{and} \quad A \cup (B \cup C) = (A \cup B) \cup C.$$

The **distributive laws** are

$$A \cap (B \cup C) \;=\; (A \cap B) \cup (A \cap C)$$

and

$$A \cup (B \cap C) \;=\; (A \cup B) \cap (A \cup C).$$

DeMorgan's laws are

$$(A \cap B)^c \;=\; A^c \cup B^c \quad \text{and} \quad (A \cup B)^c \;=\; A^c \cap B^c.$$

We next consider infinite collections of subsets of Ω. Suppose $A_n \subset \Omega$, $n = 1, 2, \ldots$. Then

$$\bigcup_{n=1}^{\infty} A_n \;:=\; \{\omega \in \Omega : \omega \in A_n \text{ for some } n \geq 1\}.$$

In other words, $\omega \in \bigcup_{n=1}^{\infty} A_n$ if and only if for at least one integer $n \geq 1$, $\omega \in A_n$. This definition admits the possibility that $\omega \in A_n$ for more than one value of n. Next, we define

$$\bigcap_{n=1}^{\infty} A_n \;:=\; \{\omega \in \Omega : \omega \in A_n \text{ for all } n \geq 1\}.$$

In other words, $\omega \in \bigcap_{n=1}^{\infty} A_n$ if and only if $\omega \in A_n$ for every positive integer n.

***Example* 1.1.** Let Ω denote the real numbers, $\Omega = (\Leftrightarrow\infty, \infty)$. Then the following infinite intersections and unions can be simplified. Consider the intersection

$$\bigcap_{n=1}^{\infty} (\Leftrightarrow\infty, 1/n) \;=\; \{\omega : \omega < 1/n, \text{for all } n \geq 1\}.$$

Now, if $\omega < 1/n$ for all $n \geq 1$, then ω cannot be positive; i.e., we must have $\omega \leq 0$. Conversely, if $\omega \leq 0$, then for all $n \geq 1$, $\omega \leq 0 < 1/n$. It follows that

$$\bigcap_{n=1}^{\infty} (\Leftrightarrow\infty, 1/n) \;=\; (\Leftrightarrow\infty, 0].$$

Consider the infinite union,

$$\bigcup_{n=1}^{\infty} (\Leftrightarrow\infty, \Leftrightarrow 1/n] \;=\; \{\omega : \omega \leq \Leftrightarrow 1/n, \text{for some } n \geq 1\}.$$

Now, if $\omega \leq \Leftrightarrow 1/n$ for some $n \geq 1$, then we must have $\omega < 0$. Conversely, if $\omega < 0$, then for large enough n, $\omega \leq \Leftrightarrow 1/n$. Thus,

$$\bigcup_{n=1}^{\infty} (\Leftrightarrow\infty, \Leftrightarrow 1/n] \;=\; (\Leftrightarrow\infty, 0).$$

In a similar way, one can show that

$$\bigcap_{n=1}^{\infty} [0, 1/n) \;=\; \{0\},$$

as well as

$$\bigcup_{n=1}^{\infty} (\Leftrightarrow\infty, n] \;=\; (\Leftrightarrow\infty, \infty) \quad \text{and} \quad \bigcap_{n=1}^{\infty} (\Leftrightarrow\infty, \Leftrightarrow n] \;=\; \emptyset.$$

The following **generalized distributive laws** also hold,

$$B \cap \left(\bigcup_{n=1}^{\infty} A_n \right) \;=\; \bigcup_{n=1}^{\infty} (B \cap A_n),$$

and

$$B \cup \left(\bigcap_{n=1}^{\infty} A_n \right) \;=\; \bigcap_{n=1}^{\infty} (B \cup A_n).$$

We also have the **generalized DeMorgan's laws**,

$$\left(\bigcap_{n=1}^{\infty} A_n \right)^c \;=\; \bigcup_{n=1}^{\infty} A_n^c,$$

and

$$\left(\bigcup_{n=1}^{\infty} A_n \right)^c \;=\; \bigcap_{n=1}^{\infty} A_n^c.$$

Finally, we will need the following definition. We say that subsets $A_n, n = 1, 2, \ldots,$ are **pairwise disjoint** if $A_n \cap A_m = \emptyset$ for all $n \neq m$.

1.2. Probability Models

Consider the experiment of tossing a fair die and measuring, i.e., noting, the face turned up. Our intuition tells us that the "probability" of the ith face turning up is $1/6$, and that the "probability" of a face with an even number of dots turning up is $1/2$.

Here is a *mathematical model* for this experiment and measurement. Let Ω be any set containing six points. We call Ω the **sample space**. Each point in Ω corresponds to, or models, a possible outcome of the experiment. For simplicity, let

$$\Omega \;:=\; \{1, 2, 3, 4, 5, 6\}.$$

Now put

$$F_i \;:=\; \{i\}, \quad i = 1, 2, 3, 4, 5, 6,$$

and

$$E := \{2, 4, 6\}.$$

We call the sets F_i and E **events**. The event F_i corresponds to, or models, the die's turning up showing the ith face. Similarly, the event E models the die's showing a face with an even number of dots. Next, for every subset A of Ω, we denote the number of points in A by $|A|$. We call $|A|$ the **cardinality** of A. We *define* the **probability** of any event A by

$$\wp(A) := |A|/|\Omega|.$$

It then follows that $\wp(F_i) = 1/6$ and $\wp(E) = 3/6 = 1/2$, which agrees with our intuition.

We now make four observations about our model:

- (i) $\wp(\emptyset) = |\emptyset|/|\Omega| = 0/|\Omega| = 0$.
- (ii) $\wp(A) \geq 0$ for every event A.
- (iii) If A and B are mutually exclusive events, i.e., $A \cap B = \emptyset$, then $\wp(A \cup B) = \wp(A) + \wp(B)$; for example, $F_3 \cap E = \emptyset$, and it is easy to check that $\wp(F_3 \cup E) = \wp(\{2, 3, 4, 6\}) = \wp(F_3) + \wp(E)$.
- (iv) When the die is tossed, *something* happens; this is modeled mathematically by the easily verified fact that $\wp(\Omega) = 1$.

As we shall see, these four properties hold for all the models discussed in this section.

We next modify our model to accommodate an unfair die as follows. Observe that for a fair die,*

$$\wp(A) = \frac{|A|}{|\Omega|} = \sum_{\omega \in A} \frac{1}{|\Omega|} = \sum_{\omega \in A} p(\omega),$$

where $p(\omega) := 1/|\Omega|$. For an *unfair* die, we redefine $\wp$ by taking

$$\wp(A) := \sum_{\omega \in A} p(\omega),$$

where now $p(\omega)$ is not constant, but is chosen to reflect the likelihood of occurrence of the various faces. This new definition of $\wp$ still satisfies (i) and (iii); however, to guarantee that (ii) and (iv) still hold, we must require that p be nonnegative and sum to one, or, in symbols, $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$.

***Example* 1.2.** Consider a die for which face i is twice as likely as face $i \Leftrightarrow 1$. Find the probability of the die's showing a face with an even number of dots.

Solution. To solve the problem, we use the above modified probability model with† $p(\omega) = 2^{\omega - 1}/63$. We then need to realize that we are being asked to compute $\wp(E)$, where $E = \{2, 4, 6\}$. Hence,

$$\wp(E) = [2^1 + 2^3 + 2^5]/63 = 42/63 = 2/3.$$

*If $A = \emptyset$, the summation is always taken to be zero.

†The derivation of this formula appears in Problem 8.

This problem is typical of the kinds of "word problems" to which probability theory is applied to analyze well-defined physical experiments. The application of probability theory requires the modeler to take the following steps:

1. Select a suitable sample space Ω.
2. Define $\wp(A)$ for all events A.
3. Translate the given "word problem" into a problem requiring the calculation of $\wp(E)$ for some specific event E.

The following example gives a family of constructions that can be used to model experiments having a finite number of possible outcomes.

***Example* 1.3.** Let M be a positive integer, and put $\Omega := \{1, 2, \ldots, M\}$. Next, let $p(1), \ldots, p(M)$ be nonnegative real numbers such that $\sum_{\omega=1}^{M} p(\omega) = 1$. For any subset $A \subset \Omega$, put

$$\wp(A) \;:=\; \sum_{\omega \in A} p(\omega).$$

In particular, to model equally likely outcomes, or equivalently, outcomes that occur "at random," we take $p(\omega) = 1/M$. In this case, $\wp(A)$ reduces to $|A|/|\Omega|$.

***Example* 1.4.** A single card is drawn at random from a well-shuffled deck of playing cards. Find the probability of drawing an ace. Also find the probability of drawing a face card.

Solution. The first step in the solution is to specify the sample space Ω and the probability $\wp$. Since there are 52 possible outcomes, we take $\Omega := \{1, \ldots, 52\}$. Each integer corresponds to one of the cards in the deck. To specify $\wp$, we must define $\wp(E)$ for all events $E \subset \Omega$. Since all cards are equally likely to be drawn, we put $\wp(E) := |E|/|\Omega|$.

To find the desired probabilities, let $1, 2, 3, 4$ correspond to the four aces, and let $41, \ldots, 52$ correspond to the 12 face cards. We identify the drawing of an ace with the event $A := \{1, 2, 3, 4\}$, and we identify the drawing of a face card with the event $F := \{41, \ldots, 52\}$. It then follows that $\wp(A) = |A|/52 = 4/52 = 1/13$ and $\wp(F) = |F|/52 = 12/52 = 3/13$.

While the sample spaces Ω in Example 1.3 can model any experiment with a finite number of outcomes, it is often convenient to use alternative sample spaces.

***Example* 1.5.** Suppose that we have two well-shuffled decks of cards, and we draw one card at random from each deck. What is the probability of drawing the ace of spades followed by the jack of hearts? What is the probability of drawing an ace and a jack (in either order)?

Solution. The first step in the solution is to specify the sample space Ω and the probability $\wp$. Since there are 52 possibilities for each draw, there are $52^2 = 2{,}704$ possible outcomes when drawing two cards. Let $D := \{1, \ldots, 52\}$, and put

$$\Omega := \{(i, j) : i, j \in D\}.$$

Then $|\Omega| = |D|^2 = 52^2 = 2{,}704$ as required. Since all pairs are equally likely, we put $\wp(E) := |E|/|\Omega|$ for arbitrary events $E \subset \Omega$.

As in the preceding example, we denote the aces by $1, 2, 3, 4$. We let 1 denote the ace of spades. We also denote the jacks by $41, 42, 43, 44$, and the jack of hearts by 42. The drawing of the ace of spades followed by the jack of hearts is identified with the event

$$A := \{(1, 42)\},$$

and so $\wp(A) = 1/2{,}704 \approx 0.000370$. The drawing of an ace and a jack is identified with $B := B_{\rm aj} \cup B_{\rm ja}$, where

$$B_{\rm aj} := \big\{(i, j) : i \in \{1, 2, 3, 4\} \text{ and } j \in \{41, 42, 43, 44\}\big\}$$

corresponds to the drawing of an ace followed by a jack, and

$$B_{\rm ja} := \big\{(i, j) : i \in \{41, 42, 43, 44\} \text{ and } j \in \{1, 2, 3, 4\}\big\}$$

corresponds to the drawing of a jack followed by an ace. Since $B_{\rm aj}$ and $B_{\rm ja}$ are disjoint, $\wp(B) = \wp(B_{\rm aj}) + \wp(B_{\rm ja}) = (|B_{\rm aj}| + |B_{\rm ja}|)/|\Omega|$. Since $|B_{\rm aj}| = |B_{\rm ja}| = 16$, $\wp(B) = 2 \cdot 16/2{,}704 = 2/169 \approx 0.0118$.

***Example* 1.6.** Two cards are drawn at random from a *single* well-shuffled deck of playing cards. What is the probability of drawing the ace of spades followed by the jack of hearts? What is the probability of drawing an ace and a jack (in either order)?

Solution. The first step in the solution is to specify the sample space Ω and the probability $\wp$. There are 52 possibilities for the first draw and 51 possibilities for the second. Hence, the sample space should contain $52 \cdot 51 = 2{,}652$ elements. Using the notation of the preceding example, we take

$$\Omega := \{(i, j) : i, j \in D \text{ with } i \neq j\},$$

Note that $|\Omega| = 52^2 \Leftrightarrow 52 = 2{,}652$ as required. Again, all such pairs are equally likely, and so we take $\wp(E) := |E|/|\Omega|$ for arbitrary events $E \subset \Omega$. The events A and B are defined as before, and the calculation is the same except that $|\Omega| = 2{,}652$ instead of 2,704. Hence, $\wp(A) = 1/2{,}652 \approx 0.000377$, and $\wp(B) = 2 \cdot 16/2{,}652 = 8/663 \approx 0.012$.

***Example* 1.7** (The Birthday Problem). In a group of n people, what is the probability that two or more people have the same birthday?

Solution. The first step in the solution is to specify the sample space Ω and the probability $\wp$. Let $D := \{1, \ldots, 365\}$ denote the days of the year, and let

$$\Omega \;:=\; \{(d_1, \ldots, d_n) : d_i \in D\}$$

denote the set of all possible sequences of n birthdays. Then $|\Omega| = |D|^n$. Since all sequences are equally likely, we take $\wp(E) := |E|/|\Omega|$ for arbitrary events $E \subset \Omega$.

Let Q denote the set of sequences $(d_1, \ldots, d_n)$ that have at least one pair of repeated entries. For example, if $n = 9$, one of the sequences in Q would be

$$(364, 17, 201, 17, 51, 171, 51, 33, 51).$$

Notice that 17 appears twice and 51 appears 3 times. The set Q is very complicated. On the other hand, consider Q^c, which is the set of sequences $(d_1, \ldots, d_n)$ that have *no* repeated entries. A typical element of Q^c can be constructed as follows. There are $|D|$ possible choices for d_1. For d_2 there are $|D| \Leftrightarrow 1$ possible choices because repetitions are not allowed for elements of Q^c. For d_3 there are $|D| \Leftrightarrow 2$ possible choices. Continuing in this way, we see that

$$|Q^c| \;=\; |D| \cdot (|D| \Leftrightarrow 1) \cdot (|D| \Leftrightarrow [n \Leftrightarrow 1]) \;=\; \frac{|D|!}{(|D| \Leftrightarrow n)!}.$$

It follows that $\wp(Q^c) = |Q^c|/|\Omega|$, and $\wp(Q) = 1 \Leftrightarrow \wp(Q^c)$. A plot of $\wp(Q)$ as a function of n is shown in Figure 1.4. As the dashed line indicates, for $n \geq 23$, the probability of two more more people having the same birthday is greater than $1/2$.

***Example* 1.8.** A certain memory location in an old personal computer is faulty and returns 8-bit bytes at random. What is the probability that a returned byte has seven 0s and one 1? Six 0s and two 1s?

Solution. Let $\Omega := \{(b_1, \ldots, b_8) : b_i = 0 \text{ or } 1\}$. Since all bytes are equally likely, put $\wp(E) := |E|/|\Omega|$ for arbitrary $E \subset \Omega$. Let

$$A_1 \;:=\; \left\{(b_1, \ldots, b_8) : \sum_{i=1}^{8} b_i = 1\right\}$$

and

$$A_2 \;:=\; \left\{(b_1, \ldots, b_8) : \sum_{i=1}^{8} b_i = 2\right\}.$$

Then $\wp(A_1) = |A_1|/|\Omega| = 8/256 = 1/32$, and $\wp(A_2) = |A_2|/|\Omega| = 28/256 = 7/64$.

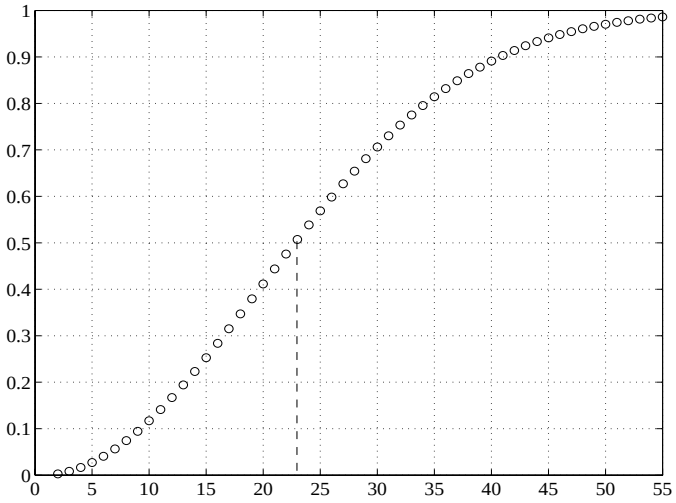

Figure 1.4. A plot of $\wp(Q)$ as a function of n. For $n \geq 23$, the probability of two or more people having the same birthday is greater than $1/2$.

In some experiments, the number of possible outcomes is countably infinite. For example, consider the tossing of a coin until the first heads appears. Here is a model for such situations. Let Ω denote the set of all positive integers, $\Omega := \{1, 2, \ldots\}$. For $\omega \in \Omega$, let $p(\omega)$ be nonnegative, and suppose that $\sum_{\omega=1}^{\infty} p(\omega) = 1$. For any subset $A \subset \Omega$, put

$$\wp(A) \;:=\; \sum_{\omega \in A} p(\omega).$$

This construction can be used to model the coin tossing experiment by identifying $\omega = i$ with the outcome that the first heads appears on the ith toss. If the probability of tails on a single toss is α $(0 \leq \alpha < 1)$, it can be shown that we should take $p(\omega) = \alpha^{\omega-1}(1 \Leftrightarrow \alpha)$ (cf. Example 2.8. To find the probability that the first head occurs before the fourth toss, we compute $\wp(A)$, where $A = \{1, 2, 3\}$. Then

$$\wp(A) \;=\; p(1) + p(2) + p(3) \;=\; (1 + \alpha + \alpha^2)(1 \Leftrightarrow \alpha).$$

If $\alpha = 1/2$, $\wp(A) = (1 + 1/2 + 1/4)/2 = 7/8$.

For some experiments, the number of possible outcomes is more than countably infinite. Examples include the lifetime of a lightbulb or a transistor, a noise voltage in a radio receiver, and the arrival time of a city bus. In these cases, $\wp$ is usually defined as an integral,

$$\wp(A) \;:=\; \int_A f(\omega)\, d\omega, \quad A \subset \Omega,$$

for some nonnegative function f. Note that f must also satisfy $\int_\Omega f(\omega)\,d\omega = 1$.

***Example* 1.9.** Consider the following model for the lifetime of a lightbulb. For the sample space we take the nonnegative half line, $\Omega := [0,\infty)$, and we put

$$\wp(A) \;:=\; \int_A f(\omega)\,d\omega\,,$$

where, for example, $f(\omega) := e^{-\omega}$. Then the probability that the lightbulb's lifetime is between 5 and 7 time units is

$$\wp([5,7]) \;=\; \int_5^7 e^{-\omega}\,d\omega \;=\; e^{-5} \Leftrightarrow e^{-7}.$$

***Example* 1.10.** A certain bus is scheduled to pick up riders at 9:15. However, it is known that the bus arrives randomly in the 20-minute interval between 9:05 and 9:25, and departs immediately after boarding waiting passengers. Find the probability that the bus arrives at or after its scheduled pick-up time.

Solution. Let $\Omega := [5,25]$, and put

$$\wp(A) \;:=\; \int_A f(\omega)\,d\omega\,.$$

Now, the term "randomly" in the problem statement is usually taken to mean that $f(\omega) \equiv$ constant. In order that $\wp(\Omega) = 1$, we must choose the constant to be $1/\text{length}(\Omega) = 1/20$. We represent the bus arriving at or after 9:15 with the event $L := [15,25]$. Then

$$\wp(L) \;=\; \int_{[15,25]} \frac{1}{20}\,d\omega \;=\; \int_{15}^{25} \frac{1}{20}\,d\omega \;=\; \frac{25 \Leftrightarrow 15}{20} \;=\; \frac{1}{2}.$$

***Example* 1.11.** A dart is thrown at random toward a circular dartboard of radius 10 cm. Assume the thrower never misses the board. Find the probability that the dart lands within 2 cm of the center.

Solution. Let $\Omega := \{(x,y) : x^2 + y^2 \le 100\}$, and for any $A \subset \Omega$, put

$$\wp(A) \;:=\; \frac{\text{area}(A)}{\text{area}(\Omega)} \;=\; \frac{\text{area}(A)}{100\pi}.$$

We then identify the event $A := \{(x,y) : x^2 + y^2 \le 4\}$ with the dart's landing within 2 cm of the center. Hence,

$$\wp(A) \;=\; \frac{4\pi}{100\pi} \;=\; 0.04.$$

1.3. Axioms and Properties of Probability

The probability models of the preceding section suggest the following axiomatic definition of probability. Given a nonempty set Ω, called the **sample space**, and a function $\wp$ defined on the subsets of Ω, we say $\wp$ is a **probability measure** if the following four axioms are satisfied:[1]

(*i*) The empty set $\emptyset$ is called the **impossible event**. The probability of the impossible event is zero; i.e., $\wp(\emptyset) = 0$.

(*ii*) Probabilities are nonnegative; i.e., for any event A, $\wp(A) \geq 0$.

(*iii*) If $A_1, A_2, \ldots$ are events that are mutually exclusive or pairwise disjoint, i.e., $A_n \cap A_m = \emptyset$ for $n \neq m$, then[‡]

$$\wp\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \wp(A_n).$$

This property is summarized by saying that the probability of the union of disjoint events is the sum of the probabilities of the individual events, or more briefly, "the probabilities of disjoint events add."

(*iv*) The entire sample space Ω is called the **sure event** or the **certain event**. Its probability is always one; i.e., $\wp(\Omega) = 1$.

We now give an interpretation of how Ω and $\wp$ model randomness. We view the sample space Ω as being the set of all possible "states of nature." First, Mother Nature chooses a state $\omega_0 \in \Omega$. We do not know which state has been chosen. We then conduct an experiment, and based on some physical measurement, we are able to determine that $\omega_0 \in A$ for some event $A \subset \Omega$. In some cases, $A = \{\omega_0\}$, that is, our measurement reveals exactly which state ω_0 was chosen by Mother Nature. (This is the case for the events F_i defined at the beginning of Section 1.2). In other cases, the set A contains ω_0 as well as other points of the sample space. (This is the case for the event E defined at the beginning of Section 1.2). In either case, we do not know before making the measurement what measurement value we will get, and so we do not know what event A Mother Nature's ω_0 will belong to. Hence, in many applications, e.g., gambling, weather prediction, computer message traffic, etc., it is useful to compute $\wp(A)$ for various events to determine which ones are most probable.

In many situations, an event A under consideration depends on some system parameter, say η, that we can select. For example, suppose A_η occurs if we correctly receive a transmitted radio message; here η could represent a voltage threshold. Hence, we can choose η to maximize $\wp(A_\eta)$.

Consequences of the Axioms

Axioms (*i*)–(*iv*) that define a probability measure have several important implications as discussed below.

[‡]See the paragraph **Finite Disjoint Unions** below and Problem 9 for further discussion regarding this axiom.

Finite Disjoint Unions. Let N be a positive integer. By taking $A_n = \emptyset$ for $n > N$ in axiom (iii), we obtain the special case (still for pairwise disjoint events)

$$\wp\Big(\bigcup_{n=1}^{N} A_n\Big) \;=\; \sum_{n=1}^{N} \wp(A_n).$$

Remark. It is not possible to go backwards and use this special case to derive axiom (iii).

***Example* 1.12.** If A is an event consisting of a finite number of sample points, say $A = \{\omega_1, \ldots, \omega_N\}$, then[2] $\wp(A) = \sum_{n=1}^{N} \wp(\{\omega_n\})$. Similarly, if A consists of a countably many sample points, say $A = \{\omega_1, \omega_2, \ldots\}$, then directly from axiom (iii), $\wp(A) = \sum_{n=1}^{\infty} \wp(\{\omega_n\})$.

Probability of a Complement. Given an event A, we can always write $\Omega = A \cup A^c$, which is a finite disjoint union. Hence, $\wp(\Omega) = \wp(A) + \wp(A^c)$. Since $\wp(\Omega) = 1$, we find that

$$\wp(A^c) \;=\; 1 \Leftrightarrow \wp(A).$$

Monotonicity. If A and B are events, then

$$A \subset B \quad \text{implies} \quad \wp(A) \leq \wp(B).$$

To see this, first note that $A \subset B$ implies

$$B \;=\; A \cup (B \cap A^c).$$

This relation is depicted in Figure 1.5, in which the disk A is a subset of the

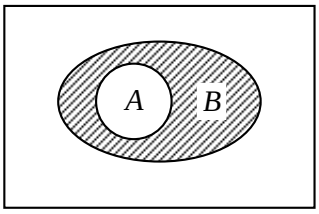

Figure 1.5. In this diagram, the disk A is a subset of the oval-shaped region B; the shaded region is $B \cap A^c$, and $B = A \cup (B \cap A^c)$.

oval-shaped region B; the shaded region is $B \cap A^c$. The figure shows that B is the disjoint union of the disk A together with the shaded region $B \cap A^c$. Since $B = A \cup (B \cap A^c)$ is a disjoint union, and since probabilities are nonnegative,

$$\begin{aligned} \wp(B) &= \wp(A) + \wp(B \cap A^c) \\ &\geq \wp(A). \end{aligned}$$

Note that the special case $B = \Omega$ results in $\wp(A) \leq 1$ for every event A. In other words, *probabilities are always less than or equal to one.*

Inclusion-Exclusion. Given any two events A and B, we always have

$$\wp(A \cup B) \;=\; \wp(A) + \wp(B) \Leftrightarrow \wp(A \cap B). \tag{1.1}$$

To derive (1.1), first note that (see Figure 1.6)

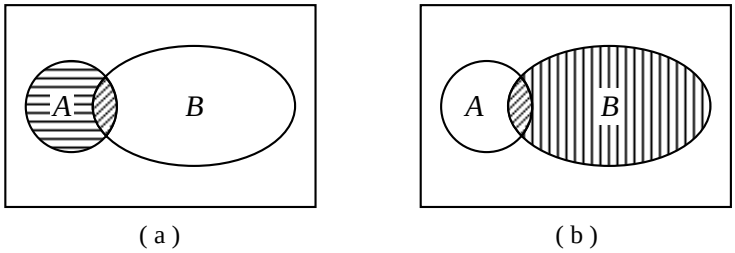

Figure 1.6. (a) Decomposition $A = (A \cap B^c) \cup (A \cap B)$. (b) Decomposition $B = (A \cap B) \cup (A^c \cap B)$.

$$A \;=\; (A \cap B^c) \cup (A \cap B)$$

and

$$B \;=\; (A \cap B) \cup (A^c \cap B).$$

Hence,

$$A \cup B \;=\; \big[(A \cap B^c) \cup (A \cap B)\big] \cup \big[(A \cap B) \cup (A^c \cap B)\big].$$

The two copies of $A \cap B$ can be reduced to one using the identity $F \cup F = F$ for any set F. Thus,

$$A \cup B \;=\; (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B).$$

A Venn diagram depicting this last decomposition is shown in Figure 1.7. Taking probabilities of the preceding equations, which involve disjoint unions, we

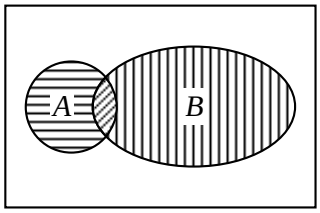

Figure 1.7. Decomposition $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$.

find that

$$\begin{aligned} \wp(A) &= \wp(A \cap B^c) + \wp(A \cap B), \\ \wp(B) &= \wp(A \cap B) + \wp(A^c \cap B), \\ \wp(A \cup B) &= \wp(A \cap B^c) + \wp(A \cap B) + \wp(A^c \cap B). \end{aligned}$$

Using the first two equations, solve for $\wp(A \cap B^c)$ and $\wp(A^c \cap B)$, respectively, and then substitute into the first and third terms on the right-hand side of the last equation. This results in

$$\begin{aligned} \wp(A \cup B) &= \big[\wp(A) \Leftrightarrow \wp(A \cap B)\big] + \wp(A \cap B) \\ &\quad + \big[\wp(B) \Leftrightarrow \wp(A \cap B)\big] \\ &= \wp(A) + \wp(B) \Leftrightarrow \wp(A \cap B). \end{aligned}$$

Limit Properties. Using axioms (i)–(iv), the following formulas can be derived (see Problems 10–12). For *any* sequence of events A_n,

$$\wp\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \lim_{N\to\infty} \wp\Big(\bigcup_{n=1}^{N} A_n\Big), \tag{1.2}$$

and

$$\wp\Big(\bigcap_{n=1}^{\infty} A_n\Big) = \lim_{N\to\infty} \wp\Big(\bigcap_{n=1}^{N} A_n\Big). \tag{1.3}$$

In particular, notice that if $A_n \subset A_{n+1}$ for all n, then the finite union in (1.2) reduces to A_N. Thus, (1.2) becomes

$$\wp\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \lim_{N\to\infty} \wp(A_N), \quad \text{if } A_n \subset A_{n+1}. \tag{1.4}$$

Similarly, if $A_{n+1} \subset A_n$ for all n, then the finite intersection in (1.3) reduces to A_N. Thus, (1.3) becomes

$$\wp\Big(\bigcap_{n=1}^{\infty} A_n\Big) = \lim_{N\to\infty} \wp(A_N), \quad \text{if } A_{n+1} \subset A_n. \tag{1.5}$$

Formulas (1.1) and (1.2) together imply that for any sequence of events A_n,

$$\wp\Big(\bigcup_{n=1}^{\infty} A_n\Big) \le \sum_{n=1}^{\infty} \wp(A_n).$$

This formula is known as the **union bound** in engineering and as **countable subadditivity** in mathematics. It is derived in Problems 13 and 14 at the end of the chapter.

1.4. Independence

Consider two tosses of an unfair coin whose "probability" of heads is 1/4. Assuming that the first toss has no influence on the second, our intuition tells us that the "probability" of heads on both tosses is $(1/4)(1/4) = 1/16$. This motivates the following definition. Two events A and B are said to be **statistically independent**, or just **independent**, if

$$\wp(A \cap B) \;=\; \wp(A)\,\wp(B). \tag{1.6}$$

The notation $A \perp\!\!\!\perp B$ is sometimes used to mean A and B are independent. We now make some important observations about independence. First, it is a simple exercise to show that if A and B are independent events, then so are A and B^c, A^c and B, and A^c and B^c. For example, using the identity

$$A \;=\; (A \cap B) \cup (A \cap B^c),$$

we have

$$\begin{aligned} \wp(A) &= \wp(A \cap B) + \wp(A \cap B^c) \\ &= \wp(A)\,\wp(B) + \wp(A \cap B^c), \end{aligned}$$

and so

$$\begin{aligned} \wp(A \cap B^c) &= \wp(A) \Leftrightarrow \wp(A)\,\wp(B) \\ &= \wp(A)[1 \Leftrightarrow \wp(B)] \\ &= \wp(A)\,\wp(B^c). \end{aligned}$$

By interchanging the roles of A and A^c and/or B and B^c, it follows that if any one of the four pairs is independent, then so are the other three.

Now suppose that A and B are any two events. If $\wp(B) = 0$, then we claim that A and B are independent. To see this, first note that the right-hand side of (1.6) is zero; as for the left-hand side, since $A \cap B \subset B$, we have $0 \le \wp(A \cap B) \le \wp(B) = 0$, i.e., the left-hand side is also zero. Hence, (1.6) always holds if $\wp(B) = 0$.

We now show that if $\wp(B) = 1$, then A and B are independent. This follows from the two preceding paragraphs. If $\wp(B) = 1$, then $\wp(B^c) = 0$, and we see that A and B^c are independent; but then A and B must also be independent.

Independence for More Than Two Events

Suppose that for $j = 1, 2, \ldots$, A_j is an event. When we say that the A_j are independent, we certainly want that for any $i \ne j$,

$$\wp(A_i \cap A_j) \;=\; \wp(A_i)\,\wp(A_j).$$

And for any distinct i, j, k, we want

$$\wp(A_i \cap A_j \cap A_k) \;=\; \wp(A_i)\,\wp(A_j)\,\wp(A_k).$$

We want analogous equations to hold for any four events, five events, and so on. In general, we want that for every *finite* subset J containing two or more positive integers,

$$\wp\Big(\bigcap_{j\in J} A_j\Big) \;=\; \prod_{j\in J} \wp(A_j).$$

In other words, we want the probability of every intersection involving finitely many of the A_j to be equal to the product of the probabilities of the individual events. If the above equation holds for all *finite* subsets of two or more positive integers, then we say that the A_j are **mutually independent**, or just independent. If the above equation holds for all subsets J containing exactly two positive integers but not necessarily for all finite subsets of 3 or more positive integers, we say that the A_j are **pairwise independent**.

***Example* 1.13.** Given three events, say A, B, and C, they are mutually independent if and only if the following equations *all* hold,

$$\begin{aligned}
\wp(A\cap B\cap C) &= \wp(A)\,\wp(B)\,\wp(C)\\
\wp(A\cap B) &= \wp(A)\,\wp(B)\\
\wp(A\cap C) &= \wp(A)\,\wp(C)\\
\wp(B\cap C) &= \wp(B)\,\wp(C).
\end{aligned}$$

It is possible to construct events A, B, and C such that the last three equations hold (pairwise independence), but the first one does not.[3] It is also possible for the first equation to hold while the last three fail.[4]

***Example* 1.14.** A coin is tossed three times and the number of heads is noted. Find the probability that the number of heads is two, assuming the tosses are mutually independent and that on each toss the probability of heads is λ for some fixed $0 \le \lambda \le 1$.

Solution. To solve the problem, let H_i denote the event that the ith toss is heads (so $\wp(H_i) = \lambda$), and let S_2 denote the event that the number of heads in three tosses is 2. Then

$$S_2 = (H_1\cap H_2\cap H_3^c)\cup(H_1\cap H_2^c\cap H_3)\cup(H_1^c\cap H_2\cap H_3).$$

This is a disjoint union, and so $\wp(S_2)$ is equal to

$$\wp(H_1\cap H_2\cap H_3^c)+\wp(H_1\cap H_2^c\cap H_3)+\wp(H_1^c\cap H_2\cap H_3). \tag{1.7}$$

Next, since H_1, H_2, and H_3 are mutually independent, so are $(H_1\cap H_2)$ and H_3. Hence, $(H_1\cap H_2)$ and H_3^c are also independent. Thus,

$$\begin{aligned}
\wp(H_1\cap H_2\cap H_3^c) &= \wp(H_1\cap H_2)\,\wp(H_3^c)\\
&= \wp(H_1)\,\wp(H_2)\,\wp(H_3^c)\\
&= \lambda^2(1\Leftrightarrow\lambda).
\end{aligned}$$

Treating the last two terms in (1.7) similarly, we have $\wp(S_2) = 3\lambda^2(1 \Leftrightarrow \lambda)$. If the coin is fair, i.e., $\lambda = 1/2$, then $\wp(S_2) = 3/8$.

In working the preceding example, we did not explicitly specify the sample space Ω or the probability measure $\wp$. This is common practice. However, the interested reader can find one possible choice for Ω and $\wp$ in the Notes.[5]

***Example* 1.15.** If $A_1, A_2, \ldots$ are mutually independent, show that

$$\wp\Big(\bigcap_{n=1}^{\infty} A_n\Big) \;=\; \prod_{n=1}^{\infty} \wp(A_n).$$

Solution. Write

$$\begin{aligned}
\wp\Big(\bigcap_{n=1}^{\infty} A_n\Big) &= \lim_{N\to\infty} \wp\Big(\bigcap_{n=1}^{N} A_n\Big), \quad \text{by (1.3)}, \\
&= \lim_{N\to\infty} \prod_{n=1}^{N} \wp(A_n), \quad \text{by independence}, \\
&= \prod_{n=1}^{\infty} \wp(A_n),
\end{aligned}$$

where the last step is just the definition of the infinite product.

***Example* 1.16.** Consider an infinite sequence of independent coin tosses. Assume that the probability of heads is $0 < p < 1$. What is the probability of seeing all heads? What is the probability of ever seeing heads?

Solution. We use the result of the preceding example as follows. Let Ω be a sample space equipped with a probability measure $\wp$ and events $A_n, n = 1, 2, \ldots$, with $\wp(A_n) = p$, where the A_n are mutually independent.[6] The event A_n corresponds to, or models, the outcome that the nth toss results in a heads. The outcome of seeing all heads corresponds to the event $\bigcap_{n=1}^{\infty} A_n$, and its probability is

$$\wp\Big(\bigcap_{n=1}^{\infty} A_n\Big) \;=\; \lim_{N\to\infty} \prod_{n=1}^{N} \wp(A_n) \;=\; \lim_{N\to\infty} p^N \;=\; 0.$$

The outcome of ever seeing heads corresponds to the event $A := \bigcup_{n=1}^{\infty} A_n$. Since $\wp(A) = 1 \Leftrightarrow \wp(A^c)$, it suffices to compute the probability of $A^c = \bigcap_{n=1}^{\infty} A_n^c$. Arguing exactly as above, we have

$$\wp\Big(\bigcap_{n=1}^{\infty} A_n^c\Big) \;=\; \lim_{N\to\infty} \prod_{n=1}^{N} \wp(A_n^c) \;=\; \lim_{N\to\infty} (1 \Leftrightarrow p)^N \;=\; 0.$$

Thus, $\wp(A) = 1 \Leftrightarrow 0 = 1$.

1.5. Conditional Probability

Conditional probability gives us a mathematically precise way of handling questions of the form, "Given that an event B has occurred, what is the *conditional probability* that A has also occurred?" If $\wp(B) > 0$, we put

$$\wp(A|B) \;:=\; \frac{\wp(A \cap B)}{\wp(B)}. \tag{1.8}$$

We call $\wp(A|B)$ the **conditional probability** of A given B. In order to justify using the word "probability" in reference to $\wp(\cdot|B)$, note that it is an easy exercise to show that if we put $Q(A) = \wp(A|B)$ for fixed B with $\wp(B) > 0$, then Q satisfies axioms (i)–(iv) in Section 1.3 that define a probability measure. Note, however, that although Q is a probability measure, it has the special property that if $A \subset B^c$, then $Q(A) = 0$. In other words, if B implies A has *not* occurred, and if B occurs, then $\wp(A|B) = 0$.

The definition of conditional probability satisfies the following intuitive property, "If A and B are independent, then $\wp(A|B)$ does not depend on B." In fact, if A and B are independent,

$$\wp(A|B) \;=\; \frac{\wp(A \cap B)}{\wp(B)} \;=\; \frac{\wp(A)\,\wp(B)}{\wp(B)} \;=\; \wp(A).$$

Observe that $\wp(A|B)$ as defined in (1.8) is the unique solution of

$$\wp(A \cap B) \;=\; \wp(A|B)\wp(B) \tag{1.9}$$

when $\wp(B) > 0$. If $\wp(B) = 0$, then no matter what real number we assign to $\wp(A|B)$, both sides of (1.9) are zero; clearly the right-hand side is zero, and, for the left-hand side, note that since $A \cap B \subset B$, we have $0 \le \wp(A \cap B) \le \wp(B) = 0$. Hence, in probability theory, when $\wp(B) = 0$, the value of $\wp(A|B)$ is permitted to be arbitrary, and it is understood that (1.9) always holds.

The Law of Total Probability and Bayes' Rule

From the identity

$$A \;=\; (A \cap B) \cup (A \cap B^c),$$

it follows that

$$\begin{aligned} \wp(A) &= \wp(A \cap B) + \wp(A \cap B^c) \\ &= \wp(A|B)\wp(B) + \wp(A|B^c)\wp(B^c). \end{aligned} \tag{1.10}$$

This formula is the simplest version of the **law of total probability**. In many applications, the quantities on the right of (1.10) are known, and it is required to find $\wp(B|A)$. This can be accomplished by writing

$$\wp(B|A) \;:=\; \frac{\wp(B \cap A)}{\wp(A)} \;=\; \frac{\wp(A \cap B)}{\wp(A)} \;=\; \frac{\wp(A|B)\wp(B)}{\wp(A)}.$$

Substituting (1.10) into the denominator yields

$$\wp(B|A) \;=\; \frac{\wp(A|B)\wp(B)}{\wp(A|B)\wp(B)+\wp(A|B^c)\wp(B^c)}.$$

This formula is the simplest version of **Bayes' rule**.

***Example* 1.17.** Polychlorinated biphenyls (PCBs) are toxic chemicals. Given that you are exposed to PCBs, suppose that your conditional probability of developing cancer is 1/3. Given that you are *not* exposed to PCBs, suppose that your conditional probability of developing cancer is 1/4. Suppose that the probability of being exposed to PCBs is 3/4. Find the probability that you were exposed to PCBs given that you do not develop cancer.

Solution. To solve this problem, we use the notation

$$E \;=\; \{\text{exposed to PCBs}\} \quad \text{and} \quad C \;=\; \{\text{develop cancer}\}.$$

With this notation, it is easy to interpret the problem as telling us that

$$\wp(C|E) \;=\; 1/3, \quad \wp(C|E^c) \;=\; 1/4, \quad \text{and} \quad \wp(E) \;=\; 3/4, \tag{1.11}$$

and asking us to find $\wp(E|C^c)$. Before solving the problem, note that the above data implies three additional equations as follows. First, recall that $\wp(E^c) = 1 \Leftrightarrow \wp(E)$. Similarly, since conditional probability is a probability as a function of its first argument, we can write $\wp(C^c|E) = 1 \Leftrightarrow \wp(C|E)$ and $\wp(C^c|E^c) = 1 \Leftrightarrow \wp(C|E^c)$. Hence,

$$\wp(C^c|E) \;=\; 2/3, \quad \wp(C^c|E^c) \;=\; 3/4, \quad \text{and} \quad \wp(E^c) \;=\; 1/4. \tag{1.12}$$

To find the desired conditional probability, we write

$$\begin{aligned}
\wp(E|C^c) &= \frac{\wp(E\cap C^c)}{\wp(C^c)} \\
&= \frac{\wp(C^c|E)\wp(E)}{\wp(C^c)} \\
&= \frac{(2/3)\,(3/4)}{\wp(C^c)} \\
&= \frac{1/2}{\wp(C^c)}.
\end{aligned}$$

To find the denominator, we use the law of total probability to write

$$\begin{aligned}
\wp(C^c) &= \wp(C^c|E)\wp(E)+\wp(C^c|E^c)\wp(E^c) \\
&= (2/3)(3/4)+(3/4)(1/4) \;=\; 11/16.
\end{aligned}$$

Hence,

$$\wp(E|C^c) \;=\; \frac{1/2}{11/16} \;=\; \frac{8}{11}.$$

In working the preceding example, we did not explicitly specify the sample space Ω or the probability measure $\wp$. As mentioned earlier, this is common practice. However, one possible choice for Ω and $\wp$ is given in the Notes.[7]

We now generalize the law of total probability. Let B_n be a sequence of pairwise disjoint events such that $\sum_n \wp(B_n) = 1$. Then for any event A,

$$\wp(A) \;=\; \sum_n \wp(A|B_n)\wp(B_n).$$

To derive this result, put $B := \bigcup_n B_n$, and observe that

$$\wp(B) \;=\; \sum_n \wp(B_n) \;=\; 1.$$

It follows that $\wp(B^c) = 0$. Next, for any event A, $A \cap B^c \subset B^c$, and so

$$0 \;\le\; \wp(A \cap B^c) \;\le\; \wp(B^c) \;=\; 0.$$

Hence, $\wp(A \cap B^c) = 0$. Writing

$$A \;=\; (A \cap B^c) \cup (A \cap B),$$

it follows that

$$\begin{aligned}
\wp(A) &= \wp(A \cap B^c) + \wp(A \cap B) \\
&= \wp(A \cap B) \\
&= \wp\left(A \cap \left[\bigcup_n B_n\right]\right) \\
&= \wp\left(\bigcup_n [A \cap B_n]\right) \\
&= \sum_n \wp(A \cap B_n). \qquad (1.13)
\end{aligned}$$

To compute the **posterior probabilities** $\wp(B_k|A)$, write

$$\wp(B_k|A) \;=\; \frac{\wp(A \cap B_k)}{\wp(A)} \;=\; \frac{\wp(A|B_k)\wp(B_k)}{\wp(A)}.$$

Applying the law of total probability to $\wp(A)$ in the denominator yields the general form of Bayes' rule,

$$\wp(B_k|A) \;=\; \frac{\wp(A|B_k)\wp(B_k)}{\sum_n \wp(A|B_n)\wp(B_n)}.$$

1.6. Notes

Notes §1.3: Axioms and Properties of Probability

***Note* 1.** When the sample space Ω is finite or countably infinite, $\wp(A)$ is usually defined for all subsets of Ω by taking

$$\wp(A) \;:=\; \sum_{\omega \in A} p(\omega)$$

for some nonnegative function p that sums to one; i.e., $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$. (It is easy to check that if $\wp$ is defined in this way, then it satisfies the axioms of a probability measure.) However, for larger sample spaces, such as when Ω is an interval of the real line, e.g., Example 1.10, it is not possible to define $\wp(A)$ for all subsets and still have $\wp$ satisfy all four axioms. (A proof of this fact can be found in advanced texts, e.g., [4, p. 45].) The way around this difficulty is to define $\wp(A)$ only for some subsets of Ω, but not all subsets of Ω. It is indeed fortunate that this can be done in such a way that $\wp(A)$ is defined for all subsets of interest that occur in practice. A set A for which $\wp(A)$ is defined is called an **event**, and the collection of all events is denoted by $\mathcal{A}$. The triple $(\Omega, \mathcal{A}, \wp)$ is called a **probability space**.

Given that $\wp(A)$ is defined only for $A \in \mathcal{A}$, in order for the probability axioms to make sense, $\mathcal{A}$ must have certain properties. First, axiom (i) requires that $\emptyset \in \mathcal{A}$. Second, axiom (iv) requires that $\Omega \in \mathcal{A}$. Third, axiom (iii) requires that if $A_1, A_2, \ldots$ are mutually exclusive events, then their union, $\bigcup_{n=1}^{\infty} A_n$, is also an event. Additionally, we need in Section 1.4 that if $A_1, A_2, \ldots$ are arbitrary events, then so is their intersection, $\bigcap_{n=1}^{\infty} A_n$. We show below that these four requirements are satisfied if we assume only that $\mathcal{A}$ has the following three properties,

(i) The empty set $\emptyset$ belongs to $\mathcal{A}$, i.e., $\emptyset \in \mathcal{A}$.

(ii) If $A \in \mathcal{A}$, then so does its complement, A^c, i.e., $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$.

(iii) If $A_1, A_2, \ldots$ belong to $\mathcal{A}$, then so does their union, $\bigcup_{n=1}^{\infty} A_n$.

A collection of subsets with these properties is called a **σ-field** or a **σ-algebra**. Of the four requirements above, the first and third obviously hold for any σ-field $\mathcal{A}$. The second requirement holds because (i) and (ii) imply that $\emptyset^c = \Omega$ must be in $\mathcal{A}$. The fourth requirement is a consequence of (iii) and DeMorgan's law.

***Note* 2.** In light of the preceding note, we see that to guarantee that $\wp(\{\omega_n\})$ is defined in Example 1.12, it is necessary to assume that the singleton sets $\{\omega_n\}$ are events, i.e., $\{\omega_n\} \in \mathcal{A}$.

Notes §1.4: Independence

***Note* 3.** Here is an example of three events that are pairwise independent, but not mutually independent. Let

$$\Omega \;:=\; \{1, 2, 3, 4, 5, 6, 7\},$$

and put $\wp(\{\omega\}) := 1/8$ for $\omega \neq 7$, and $\wp(\{7\}) := 1/4$. Take $A := \{1,2,7\}$, $B := \{3,4,7\}$, and $C := \{5,6,7\}$. Then $\wp(A) = \wp(B) = \wp(C) = 1/2$. and $\wp(A \cap B) = \wp(A \cap C) = \wp(B \cap C) = \wp(\{7\}) = 1/4$. Hence, A and B, A and C, and B and C are pairwise independent. However, since $\wp(A \cap B \cap C) = \wp(\{7\}) = 1/4$, and since $\wp(A)\,\wp(B)\,\wp(C) = 1/8$, A, B, and C are not mutually independent.

***Note* 4.** Here is an example of three events for which $\wp(A \cap B \cap C) = \wp(A)\,\wp(B)\,\wp(C)$ but no pair is independent. Let $\Omega := \{1,2,3,4\}$. Put $\wp(\{1\}) = \wp(\{2\}) = \wp(\{3\}) = p$ and $\wp(\{4\}) = q$, where $3p + q = 1$ and $0 \leq p, q \leq 1$. Put $A := \{1,4\}$, $B := \{2,4\}$, and $C := \{3,4\}$. Then the intersection of any pair is $\{4\}$, as is the intersection of all three sets. Also, $\wp(\{4\}) = q$. Since $\wp(A) = \wp(B) = \wp(C) = p+q$, we require $(p+q)^3 = q$ and $(p+q)^2 \neq q$. Solving $3p + q = 1$ and $(p+q)^3 = q$ for q reduces to solving $8q^3 + 12q^2 \Leftrightarrow 21q + 1 = 0$. Now, $q = 1$ is obviously a root, but this results in $p = 0$, which implies mutual independence. However, since $q = 1$ is a root, it is easy to verify that

$$8q^3 + 12q^2 \Leftrightarrow 21q + 1 \;=\; (q \Leftrightarrow 1)(8q^2 + 20q \Leftrightarrow 1).$$

By the quadratic formula, the desired root is $q = \left(\Leftrightarrow 5 + 3\sqrt{3}\,\right)/4$. It then follows that $p = \left(3 \Leftrightarrow \sqrt{3}\,\right)/4$ and that $p + q = \left(\Leftrightarrow 1 + \sqrt{3}\,\right)/2$. Now just observe that $(p+q)^2 \neq q$.

***Note* 5.** Here is a choice for Ω and $\wp$ for Example 1.14. Let

$$\Omega \;:=\; \{(i,j,k) : i,j,k = 0 \text{ or } 1\},$$

with 1 corresponding to heads and 0 to tails. Now put

$$\begin{aligned} H_1 &:= \{(i,j,k) : i = 1\}, \\ H_2 &:= \{(i,j,k) : j = 1\}, \\ H_3 &:= \{(i,j,k) : k = 1\}, \end{aligned}$$

and observe that

$$\begin{aligned} H_1 &= \{\ (1,0,0)\,,\ (1,0,1)\,,\ (1,1,0)\,,\ (1,1,1)\ \}, \\ H_2 &= \{\ (0,1,0)\,,\ (0,1,1)\,,\ (1,1,0)\,,\ (1,1,1)\ \}, \\ H_3 &= \{\ (0,0,1)\,,\ (0,1,1)\,,\ (1,0,1)\,,\ (1,1,1)\ \}. \end{aligned}$$

Next, let $\wp(\{(i,j,k)\}) := \lambda^{i+j+k}(1 \Leftrightarrow \lambda)^{3-(i+j+k)}$. Since

$$H_3^c \;=\; \{\ (0,0,0)\,,\ (1,0,0)\,,\ (0,1,0)\,,\ (1,1,0)\ \},$$

$H_1 \cap H_2 \cap H_3^c = \{(1,1,0)\}$. Similarly, $H_1 \cap H_2^c \cap H_3 = \{(1,0,1)\}$, and $H_1^c \cap H_2 \cap H_3 = \{(0,1,1)\}$. Hence,

$$\begin{aligned} S_2 &= \{\ (1,1,0)\,,\ (1,0,1)\,,\ (0,1,1)\ \} \\ &= \{(1,1,0)\} \cup \{(1,0,1)\} \cup \{(0,1,1)\}, \end{aligned}$$

and thus, $\wp(S_2) = 3\lambda^2(1 \Leftrightarrow \lambda)$.

***Note* 6.** To show the *existence* of a sample space and probability measure with such independent events is beyond the scope of this book. Such constructions can be found in more advanced texts such as [4, Section 36].

Notes §1.5: Conditional Probability

***Note* 7.** Here is a choice for Ω and $\wp$ for Example 1.17. Let

$$\Omega := \{(e,c) : e,c = 0 \text{ or } 1\},$$

where $e = 1$ corresponds to exposure to PCBs, and $c = 1$ corresponds to developing cancer. We then take

$$E := \{(e,c) : e = 1\} = \{ (1,0), (1,1) \},$$

and

$$C := \{(e,c) : c = 1\} = \{ (0,1), (1,1) \}.$$

It follows that

$$E^c = \{ (0,1), (0,0) \} \quad \text{and} \quad C^c = \{ (1,0), (0,0) \}.$$

Hence, $E \cap C = \{(1,1)\}$, $E \cap C^c = \{(1,0)\}$, $E^c \cap C = \{(0,1)\}$, and $E^c \cap C^c = \{(0,0)\}$.

In order to specify a suitable probability measure on Ω, we work backwards. First, *if* a measure $\wp$ on Ω exists such that (1.11) and (1.12) hold, then

$$\begin{aligned}
\wp(\{(1,1)\}) &= \wp(E \cap C) = \wp(C|E)\wp(E) = 1/4, \\
\wp(\{(0,1)\}) &= \wp(E^c \cap C) = \wp(C|E^c)\wp(E^c) = 1/16, \\
\wp(\{(1,0)\}) &= \wp(E \cap C^c) = \wp(C^c|E)\wp(E) = 1/2, \\
\wp(\{(0,0)\}) &= \wp(E^c \cap C^c) = \wp(C^c|E^c)\wp(E^c) = 3/16.
\end{aligned}$$

This suggests that we *define* $\wp$ by

$$\wp(A) := \sum_{\omega \in A} p(\omega),$$

where $p(\omega) = p(e,c)$ is given by $p(1,1) := 1/4$, $p(0,1) := 1/16$, $p(1,0) := 1/2$, and $p(0,0) := 3/16$. Starting from this definition of $\wp$, it is not hard to check that (1.11) and (1.12) hold.

1.7. Problems

Problems §1.1: Review of Set Notation

1. For real numbers $-\infty < a < b < \infty$, we use the following notation.

$$\begin{aligned}
(a,b] &:= \{x : a < x \le b\} \\
(a,b) &:= \{x : a < x < b\} \\
[a,b) &:= \{x : a \le x < b\} \\
[a,b] &:= \{x : a \le x \le b\}.
\end{aligned}$$

We also use

$$\begin{aligned}(\Leftrightarrow\infty, b] &:= \{x : x \le b\}\\ (\Leftrightarrow\infty, b) &:= \{x : x < b\}\\ (a, \infty) &:= \{x : x > a\}\\ [a, \infty) &:= \{x : x \ge a\}.\end{aligned}$$

For example, with this notation, $(0,1]^c = (\Leftrightarrow\infty, 0] \cup (1, \infty)$ and $(0,2] \cup [1,3) = (0,3)$. Now analyze

(a) $[2,3]^c$,

(b) $(1,3) \cup (2,4)$,

(c) $(1,3) \cap [2,4)$,

(d) $(3,6] \setminus (5,7)$.

2. Sketch the following subsets of the x-y plane.

(a) $B_z := \{(x,y) : x + y \le z\}$ for $z = 0, \Leftrightarrow 1, +1$.

(b) $C_z := \{(x,y) : x > 0,\ y > 0,$ and $xy \le z\}$ for $z = 1$.

(c) $H_z := \{(x,y) : x \le z\}$ for $z = 3$.

(d) $J_z := \{(x,y) : y \le z\}$ for $z = 3$.

(e) $H_z \cap J_z$ for $z = 3$.

(f) $H_z \cup J_z$ for $z = 3$.

(g) $M_z := \{(x,y) : \max(x,y) \le z\}$ for $z = 3$, where $\max(x,y)$ is the larger of x and y. For example, $\max(7,9) = 9$. Of course, $\max(9,7) = 9$ too.

(h) $N_z := \{(x,y) : \min(x,y) \le z\}$ for $z = 3$, where $\min(x,y)$ is the smaller of x and y. For example, $\min(7,9) = 7 = \min(9,7)$.

(i) $M_2 \cap N_3$.

(j) $M_4 \cap N_3$.

3. Let Ω denote the set of real numbers, $\Omega = (\Leftrightarrow\infty, \infty)$.

(a) Use the distributive law to simplify

$$[1,4] \cap \Big([0,2] \cup [3,5] \Big).$$

(b) Use DeMorgan's law to simplify $\Big([0,1] \cup [2,3] \Big)^c$.

(c) Simplify $\displaystyle\bigcap_{n=1}^{\infty} (\Leftrightarrow 1/n, 1/n)$.

(d) Simplify $\bigcap_{n=1}^{\infty} [0, 3 + 1/(2n))$.

(e) Simplify $\bigcup_{n=1}^{\infty} [5, 7 \Leftrightarrow (3n)^{-1}]$.

(f) Simplify $\bigcup_{n=1}^{\infty} [0, n]$.

Problems §1.2: Probability Models

4. A letter of the alphabet (a–z) is generated at random. Specify a sample space Ω and a probability measure $\wp$. Compute the probability that a vowel (a, e, i, o, u) is generated.

5. A collection of plastic letters, a–z, is mixed in a jar. Two letters drawn at random, one after the other. What is the probability of drawing a vowel (a, e, i, o, u) and a consonant in either order? Two vowels in any order? Specify your sample space Ω and probability $\wp$.

6. A new baby wakes up exactly once every night. The time at which the baby wakes up occurs at random between 9 pm and 7 am. If the parents go to sleep at 11 pm, what is the probability that the parents are not awakened by the baby before they would normally get up at 7 am? Specify your sample space Ω and probability $\wp$.

7. For any real or complex number $z \neq 1$ and any positive integer N, derive the **geometric series** formula

$$\sum_{k=0}^{N-1} z^k \;=\; \frac{1 \Leftrightarrow z^N}{1 \Leftrightarrow z}, \quad z \neq 1.$$

Hint: Let $S_N := 1 + z + \cdots + z^{N-1}$, and show that $S_N \Leftrightarrow z S_N = 1 \Leftrightarrow z^N$. Then solve for S_N.

Remark. If $|z| < 1$, $|z|^N \to 0$ as $N \to \infty$. Hence,

$$\sum_{k=0}^{\infty} z^k \;=\; \frac{1}{1 \Leftrightarrow z}, \quad \text{for } |z| < 1.$$

8. Let $\Omega := \{1, \ldots, 6\}$. If $p(\omega) = 2\, p(\omega \Leftrightarrow 1)$ for $\omega = 2, \ldots, 6$, and if $\sum_{\omega=1}^{6} p(\omega) = 1$, show that $p(\omega) = 2^{\omega-1}/63$. *Hint:* Use Problem 7.

Problems §1.3: Axioms and Properties of Probability

9. Suppose that instead of axiom (*iii*) of Section 1.3, we assume only that for any two disjoint events A and B, $\wp(A \cup B) = \wp(A) + \wp(B)$. Use this assumption and induction[§] on N to show that for any *finite* sequence of pairwise disjoint events $A_1, \ldots, A_N$,

$$\wp\Big(\bigcup_{n=1}^{N} A_n\Big) \;=\; \sum_{n=1}^{N} \wp(A_n).$$

Using this result for finite N, it is not possible to derive axiom (*iii*), which is the assumption needed to derive the limit results of Section 1.3.

⋆10. The purpose of this problem is to show that any countable union can be written as a union of pairwise disjoint sets. Given any sequence of sets F_n, define a new sequence by $A_1 := F_1$, and

$$A_n \;:=\; F_n \cap F_{n-1}^c \cap \cdots \cap F_1^c, \quad n \geq 2.$$

Note that the A_n are pairwise disjoint. For finite $N \geq 1$, show that

$$\bigcup_{n=1}^{N} F_n \;=\; \bigcup_{n=1}^{N} A_n.$$

Also show that

$$\bigcup_{n=1}^{\infty} F_n \;=\; \bigcup_{n=1}^{\infty} A_n.$$

⋆11. Use the preceding problem to show that for any sequence of events F_n,

$$\wp\Big(\bigcup_{n=1}^{\infty} F_n\Big) \;=\; \lim_{N\to\infty} \wp\Big(\bigcup_{n=1}^{N} F_n\Big).$$

⋆12. Use the preceding problem to show that for any sequence of events G_n,

$$\wp\Big(\bigcap_{n=1}^{\infty} G_n\Big) \;=\; \lim_{N\to\infty} \wp\Big(\bigcap_{n=1}^{N} G_n\Big).$$

13. *The Finite Union Bound.* Show that for any finite sequence of events $F_1, \ldots, F_N$,

$$\wp\Big(\bigcup_{n=1}^{N} F_n\Big) \;\leq\; \sum_{n=1}^{N} \wp(F_n).$$

Hint: Use the inclusion-exclusion formula (1.1) and induction on N. See the last footnote for information on induction.

[§] In this case, using induction on N means that you first verify the desired result for $N = 2$. Second, you assume the result is true for some arbitrary $N \geq 2$ and then prove the desired result is true for $N + 1$.

*14. *The Infinite Union Bound.* Show that for any infinite sequence of events F_n,

$$\wp\Big(\bigcup_{n=1}^{\infty} F_n\Big) \;\le\; \sum_{n=1}^{\infty} \wp(F_n).$$

Hint: Combine Problems 11 and 13.

*15. *Borel–Cantelli Lemma.* Show that if B_n is a sequence of events for which

$$\sum_{n=1}^{\infty} \wp(B_n) \;<\; \infty, \tag{1.14}$$

then

$$\wp\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} B_k\Big) \;=\; 0.$$

Hint: Let $G := \bigcap_{n=1}^{\infty} G_n$, where $G_n := \bigcup_{k=n}^{\infty} B_k$. Now use Problem 12, the union bound of the preceding problem, and the fact that (1.14) implies

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} \wp(B_n) \;=\; 0.$$

Problems §1.4: Independence

16. A certain binary communication system has a bit-error rate of 0.1; i.e., in transmitting a single bit, the probability of receiving the bit in error is 0.1. To transmit messages, a three-bit repetition code is used. In other words, to send the message **1**, 111 is transmitted, and to send the message **0**, 000 is transmitted. At the receiver, if two or more 1s are received, the decoder decides that message **1** was sent; otherwise, i.e., if two or more zeros are received, it decides that message **0** was sent. Assuming bit errors occur independently, find the probability that the decoder puts out the wrong message. *Answer:* 0.028.

17. You and your neighbor attempt to use your cordless phones at the same time. Your phones independently select one of ten channels at random to connect to the base unit. What is the probability that both phones pick the same channel?

18. A new car is equipped with dual airbags. Suppose that they fail independently with probability p. What is the probability that at least one airbag functions properly?

19. A discrete-time FIR filter is to be found satisfying certain constraints, such as energy, phase, sign changes of the coefficients, etc. FIR filters can be thought of as vectors in some finite-dimensional space. The energy constraint implies that all suitable vectors lie in some hypercube; i.e., a

square in $\mathbb{R}^2$, a cube in $\mathbb{R}^3$, etc. It is easy to generate random vectors uniformly in a hypercube. This suggests the following Monte–Carlo procedure for finding a filter that satisfies all the desired properties. Suppose we generate vectors independently in the hypercube until we find one that has all the desired properties. What is the probability of ever finding such a filter?

20. A dart is thrown at random toward a circular dartboard of radius 10 cm. Assume the thrower never misses the board. Let A_n denote the event that the dart lands within 2 cm of the center on the nth throw. Suppose that the A_n are mutually independent and that $\wp(A_n) = p$ for some $0 < p < 1$. Find the probability that the dart never lands within 2 cm of the center.

21. Each time you play the lottery, your probability of winning is p. You play the lottery n times, and plays are independent. How large should n be to make the probability of winning at least once more than 1/2? *Answer:* For $p = 1/10^6$, $n \geq 693147$.

22. Consider the sample space $\Omega = [0, 1]$ equipped with the probability measure
$$\wp(A) \;:=\; \int_A 1\, d\omega, \quad A \subset \Omega.$$
For $A = [0, 1/2]$, $B = [0, 1/4] \cup [1/2, 3/4]$, and $C = [0, 1/8] \cup [1/4, 3/8] \cup [1/2, 5/8] \cup [3/4, 7/8]$, determine whether or not A, B, and C are mutually independent.

Problems §1.5: Conditional Probability

23. The university buys workstations from two different suppliers, Moon Microsystems (MM) and Hyped–Technology (HT). On delivery, 10% of MM's workstations are defective, while 20% of HT's workstations are defective. The university buys 140 MM workstations and 60 HT workstations.

    (a) What is the probability that a workstation is from MM? From HT?

    (b) What is the probability of a defective workstation? *Answer:* 0.13.

    (c) Given that a workstation is defective, what is the probability that it came from Moon Microsystems? *Answer:* 7/13.

24. The probability that a cell in a wireless system is overloaded is 1/3. Given that it is overloaded, the probability of a blocked call is 0.3. Given that it is not overloaded, the probability of a blocked call is 0.1. Find the conditional probability that the system is overloaded given that your call is blocked. *Answer:* 0.6.

25. A certain binary communication system operates as follows. Given that a 0 is transmitted, the conditional probability that a 1 is received is ε. Given that a 1 is transmitted, the conditional probability that a 0 is

received is δ. Assume that the probability of transmitting a 0 is the same as the probability of transmitting a 1. Given that a 1 is received, find the conditional probability that a 1 was transmitted. *Hint:* Use the notation

$$T_i := \{i \text{ is transmitted}\}, \quad i = 0, 1,$$

and

$$R_j := \{j \text{ is received}\}, \quad j = 0, 1.$$

26. Professor Random has taught probability for many years. She has found that 80% of students who do the homework pass the exam, while 10% of students who don't do the homework pass the exam. If 60% of the students do the homework, what percent of students pass the exam? Of students who pass the exam, what percent did the homework? *Answer:* 12/13.

27. The Bingy 007 jet aircraft's autopilot has conditional probability 1/3 of failure given that it employs a faulty *Hexium 4* microprocessor chip. The autopilot has conditional probability 1/10 of failure given that it employs nonfaulty chip. According to the chip manufacturer, unrel, the probability of a customer's receiving a faulty *Hexium 4* chip is 1/4. Given that an autopilot failure has occurred, find the conditional probability that a faulty chip was used. Use the following notation:

$$\begin{aligned} A_F &= \{\text{autopilot fails}\} \\ C_F &= \{\text{chip is faulty}\}. \end{aligned}$$

Answer: 10/19.

*28. You have five computer chips, two of which are known to be defective.

(a) You test one of the chips; what is the probability that it is defective?

(b) Your friend tests two chips at random and reports that one is defective and one is not. Given this information, you test one of the three remaining chips at random; what is the conditional probability that the chip you test is defective?

(c) Consider the following modification of the preceding scenario. Your friend takes away two chips at random for testing; before your friend tells you the results, you test one of the three remaining chips at random; given this (lack of) information, what is the conditional probability that the chip you test is defective? Since you have not yet learned the results of your friend's tests, intuition suggests that your conditional probability should be the same as your answer to part (a). Is your intuition correct?

29. Given events A, B, and C, show that

$$\wp(A \cap B \cap C) \;=\; \wp(A|B \cap C)\,\wp(B|C)\,\wp(C).$$

Also show that

$$\wp(A \cap C|B) \;=\; \wp(A|B)\,\wp(C|B)$$

if and only if

$$\wp(A|B \cap C) \;=\; \wp(A|B).$$

In this case, A and C are **conditionally independent** given B.

CHAPTER 2

Discrete Random Variables

In most scientific and technological applications, measurements and observations are expressed as numerical quantities, e.g., thermal noise voltages, number of web site hits, round-trip times for Internet packets, engine temperatures, wind speeds, loads on an electric power grid, etc. Traditionally, numerical measurements or observations that have uncertain variability each time they are repeated are called **random variables**. However, in order to exploit the axioms and properties of probability that we studied in Chapter 1, we need to define random variables in terms of an underlying sample space Ω. Fortunately, once some basic operations on random variables are derived, we can think of random variables in the traditional manner.

The purpose of this chapter is to show how to use random variables to model events and to express probabilities and averages. Section 2.1 introduces the notions of random variable and of independence of random variables. Several examples illustrate how to express events and probabilities using multiple random variables. For discrete random variables, some of the more common probability mass functions are introduced as they arise naturally in the examples. (A summary of the more common ones can be found on the inside of the front cover.) Expectation for discrete random variables is defined in Section 2.2, and then moments and probability generating functions are introduced. Probability generating functions are an important tool for computing both probabilities and expectations. In particular, the binomial random variable arises in a natural way using generating functions, avoiding the usual combinatorial development. The weak law of large numbers is derived in Section 2.3 using Chebyshev's inequality. Conditional probability and conditional expectation are developed in Sections 2.4 and 2.5, respectively. Several examples illustrate how probabilities and expectations of interest can be easily computed using the law of total probability.

2.1. Probabilities Involving Random Variables

A real-valued function $X(\omega)$ defined for points ω in a sample space Ω is called a **random variable**.

***Example* 2.1.** Let us construct a model for counting the number of heads in a sequence of three coin tosses. For the underlying sample space, we take

$$\Omega := \{\text{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH}\},$$

which contains the eight possible sequences of tosses. However, since we are only interested in the number of heads in each sequence, we define the random

variable (function) X by

$$X(\omega) := \begin{cases} 0, & \omega = \text{TTT}, \\ 1, & \omega \in \{\text{TTH, THT, HTT}\}, \\ 2, & \omega \in \{\text{THH, HTH, HHT}\}, \\ 3, & \omega = \text{HHH}. \end{cases}$$

This is illustrated in Figure 2.1.

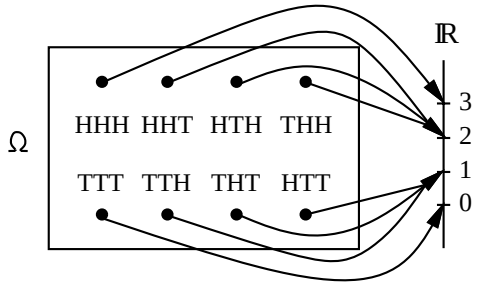

Figure 2.1. Illustration of a random variable X that counts the number of heads in a sequence of three coin tosses.

With the setup of the previous example, let us assume for specificity that the sequences are equally likely. Now let us find the probability that the number of heads X is less than 2. In other words, we want to find $\wp(X < 2)$. But what does this mean? Let us agree that $\wp(X < 2)$ is shorthand for

$$\wp(\{\omega \in \Omega : X(\omega) < 2\}).$$

Then the first step is to identify the event $\{\omega \in \Omega : X(\omega) < 2\}$. In Figure 2.1, the only lines pointing to numbers less than 2 are the lines pointing to 0 and 1. Tracing these lines backwards from $\mathbb{R}$ into Ω, we see that

$$\{\omega \in \Omega : X(\omega) < 2\} = \{\text{TTT, TTH, THT, HTT}\}.$$

Since the sequences are equally likely,

$$\begin{aligned} \wp(\{\text{TTT, TTH, THT, HTT}\}) &= \frac{|\{\text{TTT, TTH, THT, HTT}\}|}{|\Omega|} \\ &= \frac{4}{8} = \frac{1}{2}. \end{aligned}$$

The shorthand introduced above is standard in probability theory. More generally, if $B \subset \mathbb{R}$, we use the shorthand

$$\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\}$$

and[1]

$$\wp(X \in B) \;:=\; \wp(\{X \in B\}) \;=\; \wp(\{\omega \in \Omega : X(\omega) \in B\}).$$

If B is an interval such as $B = (a, b]$,

$$\{X \in (a, b]\} \;:=\; \{a < X \le b\} \;:=\; \{\omega \in \Omega : a < X(\omega) \le b\}$$

and

$$\wp(a < X \le b) \;=\; \wp(\{\omega \in \Omega : a < X(\omega) \le b\}).$$

Analogous notation applies to intervals such as $[a, b]$, $[a, b)$, (a, b), $(\Leftrightarrow\infty, b)$, $(\Leftrightarrow\infty, b]$, (a, ∞), and $[a, \infty)$.

***Example* 2.2.** Show that

$$\wp(a < X \le b) \;=\; \wp(X \le b) \Leftrightarrow \wp(X \le a).$$

Solution. It is convenient to first rewrite the desired equation as

$$\wp(X \le b) \;=\; \wp(X \le a) + \wp(a < X \le b).$$

Now observe that

$$\{\omega \in \Omega : X(\omega) \le b\} \;=\; \{\omega \in \Omega : X(\omega) \le a\} \cup \{\omega \in \Omega : a < X(\omega) \le b\}.$$

Since we cannot have an ω with $X(\omega) \le a$ and $X(\omega) > a$ at the same time, the events in the union are disjoint. Taking probabilities of both sides yields the desired result.

If B is a singleton set, say $B = \{x_0\}$, we write $\{X = x_0\}$ instead of $\big\{X \in \{x_0\}\big\}$.

***Example* 2.3.** Show that

$$\wp(X = 0 \text{ or } X = 1) \;=\; \wp(X = 0) + \wp(X = 1).$$

Solution. First note that the word "or" means "union." Hence, we are trying to find the probability of $\{X = 0\} \cup \{X = 1\}$. If we expand our shorthand, this union becomes

$$\{\omega \in \Omega : X(\omega) = 0\} \cup \{\omega \in \Omega : X(\omega) = 1\}.$$

Since we cannot have an ω with $X(\omega) = 0$ and $X(\omega) = 1$ at the same time, these events are disjoint. Hence, their probabilities add, and we obtain

$$\wp(\{X = 0\} \cup \{X = 1\}) \;=\; \wp(X = 0) + \wp(X = 1).$$

The following notation is also useful in computing probabilities. For any subset B, we define the **indicator function** of B by

$$I_B(x) \;:=\; \begin{cases} 1, & x \in B, \\ 0, & x \notin B. \end{cases}$$

Readers familiar with the **unit-step function**,

$$u(x) \;:=\; \begin{cases} 1, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

will note that $u(x) = I_{[0,\infty)}(x)$. However, the indicator notation is often more compact. For example, if $a < b$, it is easier to write $I_{[a,b)}(x)$ than $u(x \Leftrightarrow a) \Leftrightarrow u(x \Leftrightarrow b)$. How would you write $I_{(a,b]}(x)$ in terms of the unit step?

Discrete Random Variables

We say X is a **discrete random variable** if there exist distinct real numbers x_i such that

$$\sum_i \wp(X = x_i) \;=\; 1.$$

For discrete random variables,

$$\wp(X \in B) \;=\; \sum_i I_B(x_i)\wp(X = x_i).$$

To derive this equation, we apply the law of total probability as given in (1.13) with $A = \{X \in B\}$ and $B_i = \{X = x_i\}$. Since the x_i are distinct, the B_i are disjoint, and (1.13) says that

$$\wp(X \in B) \;=\; \sum_i \wp(\{X \in B\} \cap \{X = x_i\}).$$

Now observe that

$$\{X \in B\} \cap \{X = x_i\} \;=\; \begin{cases} \{X = x_i\}, & x_i \in B, \\ \emptyset, & x_i \notin B. \end{cases}$$

Hence,

$$\wp(\{X \in B\} \cap \{X = x_i\}) \;=\; I_B(x_i)\wp(X = x_i).$$

Integer-Valued Random Variables

An **integer-valued random variable** is a discrete random variable whose distinct values are $x_i = i$. For integer-valued random variables,

$$\wp(X \in B) \;=\; \sum_i I_B(i)\wp(X = i).$$

Here are some simple probability calculations involving integer-valued random variables.

$$\wp(X \le 7) \;=\; \sum_i I_{(-\infty,7]}(i)\wp(X=i) \;=\; \sum_{i=-\infty}^{7} \wp(X=i).$$

Similarly,

$$\wp(X \ge 7) \;=\; \sum_i I_{[7,\infty)}(i)\wp(X=i) \;=\; \sum_{i=7}^{\infty} \wp(X=i).$$

However,

$$\wp(X > 7) \;=\; \sum_i I_{(7,\infty)}(i)\wp(X=i) \;=\; \sum_{i=8}^{\infty} \wp(X=i),$$

which is equal to $\wp(X \ge 8)$. Similarly

$$\wp(X < 7) \;=\; \sum_i I_{(-\infty,7)}(i)\wp(X=i) \;=\; \sum_{i=-\infty}^{6} \wp(X=i),$$

which is equal to $\wp(X \le 6)$.

When an experiment results in a finite number of "equally likely" or "totally random" outcomes, we model it with a **uniform** random variable. We say that X is uniformly distributed on $1, \ldots, n$ if

$$\wp(X=k) \;=\; \frac{1}{n}, \quad k = 1, \ldots, n.$$

For example, to model the toss of a fair die we would use $\wp(X=k) = 1/6$ for $k = 1, \ldots, 6$. To model the selection of one card for a well-shuffled deck of playing cards we would use $\wp(X=k) = 1/52$ for $k = 1, \ldots, 52$. More generally, we can let k vary over any subset of n integers. A common alternative to $1, \ldots, n$ is $0, \ldots, n \Leftrightarrow 1$. For k not in the range of experimental outcomes, we put $\wp(X=k) = 0$.

***Example* 2.4.** Ten neighbors each have a cordless phone. The number of people using their cordless phones at the same time is totally random. Find the probability that more than half of the phones are in use at the same time.

Solution. We model the number of phones in use at the same time as a uniformly distributed random variable X taking values $0, \ldots, 10$. Zero is included because we allow for the possibility that no phones are in use. We must compute

$$\wp(X > 5) \;=\; \sum_{i=6}^{10} \wp(X=i) \;=\; \sum_{i=6}^{10} \frac{1}{11} \;=\; \frac{5}{11}.$$

If the preceding example had asked for the probability that at least half the phones are in use, then the answer would have been $\wp(X \ge 5) = 6/11$.

Pairs of Random Variables

If X and Y are random variables, we use the shorthand

$$\{X \in B, Y \in C\} \;:=\; \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\},$$

which is equal to

$$\{\omega \in \Omega : X(\omega) \in B\} \cap \{\omega \in \Omega : Y(\omega) \in C\}.$$

Putting all of our shorthand together, we can write

$$\{X \in B, Y \in C\} \;=\; \{X \in B\} \cap \{Y \in C\}.$$

We also have

$$\begin{aligned} \wp(X \in B, Y \in C) &:= \wp(\{X \in B, Y \in C\}) \\ &= \wp(\{X \in B\} \cap \{Y \in C\}). \end{aligned}$$

In particular, if the events $\{X \in B\}$ and $\{Y \in C\}$ are independent for all sets B and C, we say that X and Y are **independent random variables**. In light of this definition and the above shorthand, we see that X and Y are independent random variables if and only if

$$\wp(X \in B, Y \in C) \;=\; \wp(X \in B)\, \wp(Y \in C) \tag{2.1}$$

for all sets[2] B and C.

***Example* 2.5.** On a certain aircraft, the main control circuit on an autopilot fails with probability p. A redundant backup circuit fails independently with probability q. The airplane can fly if at least one of the circuits is functioning. Find the probability that the airplane cannot fly.

Solution. We introduce two random variables, X and Y. We set $X = 1$ if the main circuit fails, and $X = 0$ otherwise. We set $Y = 1$ if the backup circuit fails, and $Y = 0$ otherwise. Then $\wp(X = 1) = p$ and $\wp(Y = 1) = q$. We assume X and Y are independent random variables. Then the event that the airplane cannot fly is modeled by

$$\{X = 1\} \cap \{Y = 1\}.$$

Using the independence of X and Y, $\wp(X = 1, Y = 1) = \wp(X = 1)\wp(Y = 1) = pq$.

The random variables X and Y of the preceding example are said to be **Bernoulli**. To indicate the relevant parameters, we write $X \sim \text{Bernoulli}(p)$ and $Y \sim \text{Bernoulli}(q)$. Bernoulli random variables are good for modeling the result of an experiment having two possible outcomes (numerically represented by 0 and 1), e.g., a coin toss, testing whether a certain block on a computer disk is bad, whether a new radar system detects a stealth aircraft, whether a certain internet packet is dropped due to congestion at a router, etc.

Multiple Independent Random Variables

We say that $X_1, X_2, \ldots$ are independent random variables, if for every finite subset J containing two or more positive integers,

$$\wp\Big(\bigcap_{j\in J}\{X_j \in B_j\}\Big) \;=\; \prod_{j\in J}\wp(X_j \in B_j),$$

for all sets B_j. If for every B, $\wp(X_j \in B)$ does not depend on j for all j, then we say the X_j are **identically distributed**. If the X_j are both independent and identically distributed, we say they are **i.i.d.**

***Example* 2.6.** Ten neighbors each have a cordless phone. The number of people using their cordless phones at the same time is totally random. For one week (five days), each morning at 9:00 am we count the number of people using their cordless phone. Assume that the numbers of phones in use on different days are independent. Find the probability that on each day fewer than three phones are in use at 9:00 am. Also find the probability that on at least one day, more than two phones are in use at 9:00 am.

Solution. For $i = 1, \ldots, 5$, let X_i denote the number of phones in use at 9:00 am on the ith day. We assume that the X_i are independent, uniformly distributed random variables taking values $0, \ldots, 10$. For the first part of the problem, we must compute

$$\wp\Big(\bigcap_{i=1}^{5}\{X_i < 3\}\Big).$$

Using independence,

$$\wp\Big(\bigcap_{i=1}^{5}\{X_i < 3\}\Big) \;=\; \prod_{i=1}^{5}\wp(X_i < 3).$$

Now observe that since the X_i are nonnegative, integer-valued random variables,

$$\begin{aligned}
\wp(X_i < 3) &= \wp(X_i \le 2)\\
&= \wp(X_i = 0 \text{ or } X_i = 1 \text{ or } X_i = 2)\\
&= \wp(X_i = 0) + \wp(X_i = 1) + \wp(X_i = 2)\\
&= 3/11.
\end{aligned}$$

Hence,

$$\wp\Big(\bigcap_{i=1}^{5}\{X_i < 3\}\Big) \;=\; (3/11)^5 \;\approx\; 0.00151.$$

For the second part of the problem, we must compute

$$\wp\Big(\bigcup_{i=1}^{5}\{X_i > 2\}\Big) \;=\; 1 \Leftrightarrow \wp\Big(\bigcap_{i=1}^{5}\{X_i \le 2\}\Big)$$

$$\begin{aligned} &= 1 \Leftrightarrow \prod_{i=1}^{5} \wp(X_i \le 2) \\ &= 1 \Leftrightarrow (3/11)^5 \approx 0.99849. \end{aligned}$$

Calculations similar to those in the preceding example can be used to find probabilities involving the maximum or minimum of several independent random variables.

***Example* 2.7.** Let $X_1, \ldots, X_n$ be independent random variables. Evaluate

$$\wp(\max(X_1, \ldots, X_n) \le z) \quad \text{and} \quad \wp(\min(X_1, \ldots, X_n) \le z).$$

Solution. Observe that $\max(X_1, \ldots, X_n) \le z$ if and only if all of the X_k are less than or equal to z; i.e.,

$$\{\max(X_1, \ldots, X_n) \le z\} = \bigcap_{k=1}^{n} \{X_k \le z\}.$$

It then follows that

$$\begin{aligned} \wp(\max(X_1, \ldots, X_n) \le z) &= \wp\left(\bigcap_{k=1}^{n} \{X_k \le z\} \right) \\ &= \prod_{k=1}^{n} \wp(X_k \le z), \end{aligned}$$

where the second equation follows by independence.

For the min problem, observe that $\min(X_1, \ldots, X_n) \le z$ if and only if at least one of the X_i is less than or equal to z; i.e.,

$$\{\min(X_1, \ldots, X_n) \le z\} = \bigcup_{k=1}^{n} \{X_k \le z\}.$$

Hence,

$$\begin{aligned} \wp(\min(X_1, \ldots, X_n) \le z) &= \wp\left(\bigcup_{k=1}^{n} \{X_k \le z\} \right) \\ &= 1 \Leftrightarrow \wp\left(\bigcap_{k=1}^{n} \{X_k > z\} \right) \\ &= 1 \Leftrightarrow \prod_{k=1}^{n} \wp(X_k > z). \end{aligned}$$

***Example* 2.8.** A drug company has developed a new vaccine, and would like to analyze how effective it is. Suppose the vaccine is tested in several volunteers until one is found in whom the vaccine fails. Let $T = k$ if the *first* time the vaccine fails is on the kth volunteer. Find $\wp(T = k)$.

Solution. For $i = 1, 2, \ldots$, let $X_i = 1$ if the vaccine works in the ith volunteer, and let $X_i = 0$ if it fails in the ith volunteer. Then the first failure occurs on the kth volunteer if and only if the vaccine works for the first $k \Leftrightarrow 1$ volunteers and then fails for the kth volunteer. In terms of events,

$$\{T = k\} \;=\; \{X_1 = 1\} \cap \cdots \cap \{X_{k-1} = 1\} \cap \{X_k = 0\}.$$

Assuming the X_i are independent, and taking probabilities of both sides yields

$$\begin{aligned} \wp(T = k) &= \wp(\{X_1 = 1\} \cap \cdots \cap \{X_{k-1} = 1\} \cap \{X_k = 0\}) \\ &= \wp(X_1 = 1) \cdots \wp(X_{k-1} = 1) \cdot \wp(X_k = 0). \end{aligned}$$

If we further assume that the X_i are identically distributed with $p := \wp(X_i = 1)$ for all i, then

$$\wp(T = k) \;=\; p^{k-1}(1 \Leftrightarrow p).$$

The preceding random variable T is an example of a **geometric** random variable. Since we consider two types of geometric random variables, both depending on $0 \le p < 1$, it is convenient to write $X \sim \text{geometric}_1(p)$ if

$$\wp(X = k) \;=\; (1 \Leftrightarrow p)p^{k-1}, \quad k = 1, 2, \ldots,$$

and $X \sim \text{geometric}_0(p)$ if

$$\wp(X = k) \;=\; (1 \Leftrightarrow p)p^{k}, \quad k = 0, 1, \ldots.$$

As the above exmaple shows, the geometric$_1(p)$ random variable represents the first time something happens. We will see in Chapter 9 that the geometric$_0(p)$ random variable represents the steady-state number of customers in a queue with an infinite buffer.

Note that by the geometric series formula (Problem 7 in Chapter 1), for any real or complex number z with $|z| < 1$,

$$\sum_{k=0}^{\infty} z^k \;=\; \frac{1}{1 \Leftrightarrow z}.$$

Taking $z = p$ shows that the probabilities sum to one in both cases. Note that if we put $q = 1 \Leftrightarrow p$, then $0 < q \le 1$, and we can write $\wp(X = k) = q(1 \Leftrightarrow q)^{k-1}$ in the geometric$_1(p)$ case and $\wp(X = k) = q(1 \Leftrightarrow q)^k$ in the geometric$_0(p)$ case.

***Example* 2.9.** If $X \sim \text{geometric}_0(p)$, evaluate $\wp(X > 2)$.

Solution. We could write

$$\wp(X > 2) \;=\; \sum_{k=3}^{\infty} \wp(X = k).$$

However, since $\wp(X = k) = 0$ for negative k, a finite series is obtained by writing

$$\begin{aligned} \wp(X > 2) &= 1 \Leftrightarrow \wp(X \le 2) \\ &= 1 \Leftrightarrow \sum_{k=0}^{2} \wp(X = k) \\ &= 1 \Leftrightarrow (1 \Leftrightarrow p)[1 + p + p^2]. \end{aligned}$$

Probability Mass Functions

The **probability mass function** (pmf) of a discrete random variable X taking distinct values x_i is defined by

$$p_X(x_i) \;:=\; \wp(X = x_i).$$

With this notation

$$\wp(X \in B) \;=\; \sum_i I_B(x_i)\wp(X = x_i) \;=\; \sum_i I_B(x_i) p_X(x_i).$$

***Example* 2.10.** If $X \sim \text{geometric}_0(p)$, write down its pmf.

Solution. Write

$$p_X(k) \;:=\; \wp(X = k) \;=\; (1 \Leftrightarrow p)p^k, \quad k = 0, 1, 2, \ldots.$$

A graph of $p_X(k)$ is shown in Figure 2.2.

The **Poisson** random variable is used to model many different physical phenomena ranging from the photoelectric effect and radioactive decay to computer message traffic arriving at a queue for transmission. A random variable X is said to have a Poisson probability mass function with parameter $\lambda > 0$, denoted by $X \sim \text{Poisson}(\lambda)$, if

$$p_X(k) \;=\; \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots.$$

A graph of $p_X(k)$ is shown in Figure 2.3. To see that these probabilities sum to one, recall that for any real or complex number z, the power series for e^z is

$$e^z \;=\; \sum_{k=0}^{\infty} \frac{z^k}{k!}.$$

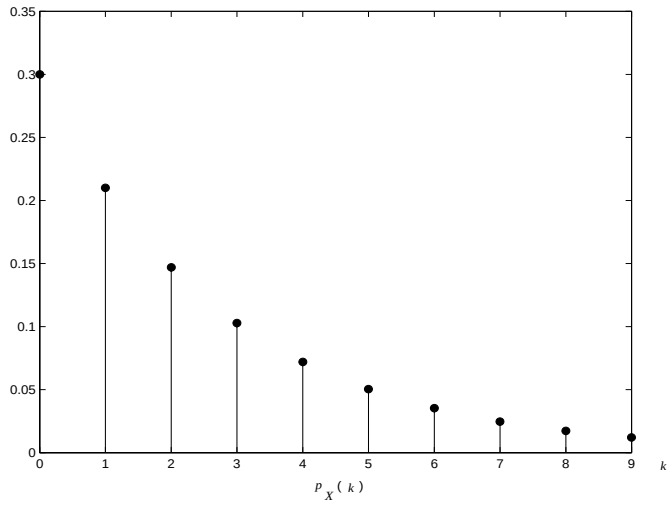

Figure 2.2. The geometric$_0(p)$ pmf $p_X(k) = (1-p)p^k$ with $p = 0.7$.

The **joint probability mass function** of X and Y is defined by

$$p_{XY}(x_i, y_j) \;:=\; \wp(X = x_i, Y = y_j) \;=\; \wp(\{X = x_i\} \cap \{Y = y_j\}). \tag{2.2}$$

Two applications of the law of total probability as in (1.13) can be used to show that[3]

$$\wp(X \in B, Y \in C) \;=\; \sum_i \sum_j I_B(x_i) I_C(y_j) p_{XY}(x_i, y_j). \tag{2.3}$$

If $B = \{x_k\}$ and $C = \mathbb{R}$, the left-hand side of (2.3) is*

$$\wp(X = x_k, Y \in \mathbb{R}) \;=\; \wp(\{X = x_k\} \cap \Omega) \;=\; \wp(X = x_k).$$

The right-hand side of (2.3) reduces to $\sum_j p_{XY}(x_k, y_j)$. Hence,

$$p_X(x_k) \;=\; \sum_j p_{XY}(x_k, y_j).$$

Thus, the pmf of X can be recovered from the joint pmf of X and Y by summing over all values of Y. Similarly,

$$p_Y(y_\ell) \;=\; \sum_i p_{XY}(x_i, y_\ell).$$

*Because we have defined random variables to be *real-valued* functions of ω, it is easy to see, after expanding our shorthand notation, that

$$\{Y \in \mathbb{R}\} \;:=\; \{\omega \in \Omega : Y(\omega) \in \mathbb{R}\} \;=\; \Omega.$$

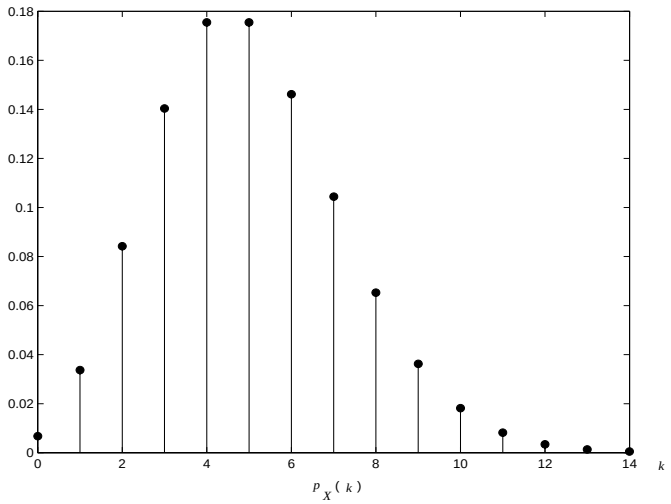

Figure 2.3. The Poisson(λ) pmf $p_X(k) = \lambda^k e^{-\lambda}/k!$ with $\lambda = 5$.

In this context, we call p_X and p_Y **marginal probability mass functions**.

From (2.2) it is clear that if X and Y are independent, then $p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j)$. The converse is also true, since if $p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j)$, then the double sum in (2.3) factors, and we have

$$\begin{aligned} \wp(X \in B, Y \in C) &= \left[\sum_i I_B(x_i) p_X(x_i)\right]\left[\sum_j I_C(y_j) p_Y(y_j)\right] \\ &= \wp(X \in B)\,\wp(Y \in C). \end{aligned}$$

Hence, for a pair of discrete random variables X and Y, they are independent if and only if their joint pmf factors into the product of their marginal pmfs.

2.2. Expectation

The notion of expectation is motivated by the conventional definition of numerical average. Recall that the numerical average of n numbers, $x_1, \ldots, x_n$, is

$$\frac{1}{n}\sum_{i=1}^{n} x_i.$$

We use the average to summarize or characterize the entire collection of numbers $x_1, \ldots, x_n$ with a single "typical" value.

If X is a discrete random variable taking n distinct, equally likely, real values, $x_1, \ldots, x_n$, then we define its expectation by

$$\mathsf{E}[X] \;:=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Since $\wp(X = x_i) = 1/n$ for each $i = 1, \ldots, n$, we can also write

$$\mathsf{E}[X] \;=\; \sum_{i=1}^{n} x_i \, \wp(X = x_i).$$

Observe that this last formula makes sense even if the values x_i do not occur with the same probability. We therefore define the **expectation**, **mean**, or **average** of a discrete random variable X by

$$\mathsf{E}[X] \;:=\; \sum_{i} x_i \, \wp(X = x_i),$$

or, using the pmf notation $p_X(x_i) = \wp(X = x_i)$,

$$\mathsf{E}[X] \;=\; \sum_{i} x_i \, p_X(x_i).$$

In complicated problems, it is sometimes easier to compute means than probabilities. Hence, the single number $\mathsf{E}[X]$ often serves as a conveniently computable way to summarize or partially characterize an entire collection of pairs $\{(x_i, p_X(x_i))\}$.

***Example* 2.11.** Find the mean of a Bernoulli(p) random variable X.

Solution. Since X takes only the values $x_0 = 0$ and $x_1 = 1$, we can write

$$\mathsf{E}[X] \;=\; \sum_{i=0}^{1} i \, \wp(X = i) \;=\; 0 \cdot (1 \Leftrightarrow p) + 1 \cdot p \;=\; p.$$

Note that while X takes only the values 0 and 1, its "typical" value p is never seen (unless $p = 0$ or $p = 1$).

***Example* 2.12** (Every Probability is an Expectation). There is a close connection between expectation and probability. If X is any random variable and B is any set, then $I_B(X)$ is a discrete random variable taking the values zero and one. Thus, $I_B(X)$ is a Bernoulli random variable, and

$$\mathsf{E}[I_B(X)] \;=\; \wp\big(I_B(X) = 1\big) \;=\; \wp(X \in B).$$

***Example* 2.13.** When light of intensity λ is incident on a photodetector, the number of photoelectrons generated is Poisson with parameter λ. Find the mean number of photoelectrons generated.

Solution. Let X denote the number of photoelectrons generated. We need to calculate $\mathsf{E}[X]$. Since a Poisson random variable takes only nonnegative integer values with positive probability,

$$\mathsf{E}[X] \;=\; \sum_{n=0}^{\infty} n \, \wp(X = n).$$

Since the term with $n = 0$ is zero, it can be dropped. Hence,

$$\begin{aligned} \mathsf{E}[X] &= \sum_{n=1}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{(n \Leftrightarrow 1)!} \\ &= \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n \Leftrightarrow 1)!}. \end{aligned}$$

Now change the index of summation from n to $k = n \Leftrightarrow 1$. This results in

$$\mathsf{E}[X] = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

***Example* 2.14** (Infinite Expectation). Here is random variable for which $\mathsf{E}[X] = \infty$. Suppose that $\wp(X = k) = C^{-1}/k^2$, $k = 1, 2, \ldots$, where[†]

$$C := \sum_{k=1}^{\infty} \frac{1}{k^2}.$$

Then

$$\mathsf{E}[X] = \sum_{k=1}^{\infty} k \, \wp(X = k) = \sum_{k=1}^{\infty} k \frac{C^{-1}}{k^2} = C^{-1} \sum_{k=1}^{\infty} \frac{1}{k} = \infty$$

as shown in Problem 34.

Some care is necessary when computing expectations of signed random variables that take more than finitely many values. It is the convention in probability theory that $\mathsf{E}[X]$ should be evaluated as

$$\mathsf{E}[X] = \sum_{i : x_i \geq 0} x_i \, \wp(X = x_i) + \sum_{i : x_i < 0} x_i \, \wp(X = x_i),$$

assuming that at least one of these sums is finite. If the first sum is $+\infty$ and the second one is $\Leftrightarrow\infty$, then no value is assigned to $\mathsf{E}[X]$, and we say that $\mathsf{E}[X]$ is undefined.

[†]Note that C is finite by Problem 34. This is important since if $C = \infty$, then $C^{-1} = 0$ and the probabilities would sum to zero instead of one.

***Example* 2.15** (Undefined Expectation). With C as in the previous example, suppose that for $k = 1, 2, \ldots$, $\wp(X = k) = \wp(X = \Leftrightarrow k) = \frac{1}{2}C^{-1}/k^2$. Then

$$\begin{aligned} \mathsf{E}[X] &= \sum_{k=1}^{\infty} k\,\wp(X = k) + \sum_{k=-\infty}^{-1} k\,\wp(X = k) \\ &= \frac{1}{2C}\sum_{k=1}^{\infty}\frac{1}{k} + \frac{1}{2C}\sum_{k=-\infty}^{-1}\frac{1}{k} \\ &= \frac{\infty}{2C} + \frac{\Leftrightarrow\infty}{2C} \\ &= \text{undefined.} \end{aligned}$$

Expectation of Functions of Random Variables, or the Law of the Unconscious Statistician (LOTUS)

Given a random variable X, we will often have occasion to define a new random variable by $Z := g(X)$, where $g(x)$ is a real-valued function of the real variable x. More precisely, recall that a random variable X is actually a function taking points of the sample space, $\omega \in \Omega$, into real numbers $X(\omega)$. Hence, the notation $Z = g(X)$ is actually shorthand for $Z(\omega) := g(X(\omega))$. We now show that if X has pmf p_X, then we can compute $\mathsf{E}[Z] = \mathsf{E}[g(X)]$ without actually finding the pmf of Z. The precise formula is

$$\mathsf{E}[g(X)] = \sum_i g(x_i)\,p_X(x_i). \tag{2.4}$$

Equation (2.4) is sometimes called the **law of the unconscious statistician** (LOTUS). As a simple example of its use, we can write, for any constant a,

$$\mathsf{E}[aX] = \sum_i a x_i\,p_X(x_i) = a\sum_i x_i\,p_X(x_i) = a\mathsf{E}[X].$$

In other words, constant factors can be pulled out of the expectation.

** Derivation of LOTUS*

To derive (2.4), we proceed as follows. Let X take distinct values x_i. Then Z takes values $g(x_i)$. However, the values $g(x_i)$ may not be distinct. For example, if $g(x) = x^2$, and X takes the four distinct values ± 1 and ± 2, then Z takes only the two distinct values 1 and 4. In any case, let z_k denote the distinct values of Z, and put

$$B_k := \{x : g(x) = z_k\}.$$

Then $Z = z_k$ if and only if $g(X) = z_k$, and this happens if and only if $X \in B_k$. Thus, the pmf of Z is

$$p_Z(z_k) := \wp(Z = z_k) = \wp(X \in B_k),$$

where

$$\wp(X \in B_k) \;=\; \sum_i I_{B_k}(x_i)\, p_X(x_i).$$

We now compute

$$\begin{aligned} \mathsf{E}[Z] &= \sum_k z_k\, p_Z(z_k) \\ &= \sum_k z_k \left[\sum_i I_{B_k}(x_i)\, p_X(x_i) \right] \\ &= \sum_i \left[\sum_k z_k I_{B_k}(x_i) \right] p_X(x_i). \end{aligned}$$

From the definition of the B_k in terms of the distinct z_k, the B_k are disjoint. Hence, for fixed i in the outer sum, in the inner sum, only one of the B_k can contain x_i. In other words, all terms but one in the inner sum are zero. Furthermore, for that one value of k with $I_{B_k}(x_i) = 1$, we have $x_i \in B_k$, and $z_k = g(x_i)$; i.e., for this value of k,

$$z_k I_{B_k}(x_i) \;=\; g(x_i) I_{B_k}(x_i) \;=\; g(x_i).$$

Hence, the inner sum reduces to $g(x_i)$, and

$$\mathsf{E}[Z] \;=\; \mathsf{E}[g(X)] \;=\; \sum_i g(x_i)\, p_X(x_i).$$

Linearity of Expectation

The derivation of the law of the unconscious statistician can be generalized to show that if $g(x, y)$ is a real-valued function of two variables x and y, then

$$\mathsf{E}[g(X,Y)] \;=\; \sum_i \sum_j g(x_i, y_j)\, p_{XY}(x_i, y_j).$$

In particular, taking $g(x, y) = x + y$, it is a simple exercise to show that $\mathsf{E}[X + Y] = \mathsf{E}[X] + \mathsf{E}[Y]$. Using this fact, we can write

$$\mathsf{E}[aX + bY] \;=\; \mathsf{E}[aX] + \mathsf{E}[bY] \;=\; a\mathsf{E}[X] + b\mathsf{E}[Y].$$

In other words, expectation is **linear**.

***Example* 2.16.** A binary communication link has bit-error probability p. What is the expected number of bit errors in a transmission of n bits?

Solution. For $i = 1, \ldots, n$, let $X_i = 1$ if the ith bit is received incorrectly, and let $X_i = 0$ otherwise. Then $X_i \sim \text{Bernoulli}(p)$, and $Y := X_1 + \cdots + X_n$ is the total number of errors in the transmission of n bits. We know from Example 2.11 that $\mathsf{E}[X_i] = p$. Hence,

$$\mathsf{E}[Y] \;=\; \mathsf{E}\left[\sum_{i=1}^{n} X_i \right] \;=\; \sum_{i=1}^{n} \mathsf{E}[X_i] \;=\; \sum_{i=1}^{n} p \;=\; np.$$

Moments

The nth **moment**, $n \geq 1$, of a real-valued random variable X is defined to be $\mathsf{E}[X^n]$. In particular, the first moment of X is its mean $\mathsf{E}[X]$. Letting $m = \mathsf{E}[X]$, we define the **variance** of X by

$$\begin{aligned}
\mathsf{var}(X) &:= \mathsf{E}[(X - m)^2] \\
&= \mathsf{E}[X^2 - 2mX + m^2] \\
&= \mathsf{E}[X^2] - 2m\mathsf{E}[X] + m^2, \quad \text{by linearity,} \\
&= \mathsf{E}[X^2] - m^2 \\
&= \mathsf{E}[X^2] - (\mathsf{E}[X])^2.
\end{aligned}$$

The variance is a measure of the spread of a random variable about its mean value. The larger the variance, the more likely we are to see values of X that are far from the mean m. The **standard deviation** of X is defined to be the positive square root of the variance. The nth **central moment** of X is $\mathsf{E}[(X - m)^n]$. Hence, the second central moment is the variance.

***Example* 2.17.** Find the second moment and the variance of X if $X \sim$ Bernoulli(p).

Solution. Since X takes only the values 0 and 1, it has the unusual property that $X^2 = X$. Hence, $\mathsf{E}[X^2] = \mathsf{E}[X] = p$. It now follows that

$$\mathsf{var}(X) \;=\; \mathsf{E}[X^2] - (\mathsf{E}[X])^2 \;=\; p - p^2 \;=\; p(1-p).$$

***Example* 2.18.** Find the second moment and the variance of X if $X \sim$ Poisson(λ).

Solution. Observe that $\mathsf{E}[X(X-1)] + \mathsf{E}[X] = \mathsf{E}[X^2]$. Since we know that $\mathsf{E}[X] = \lambda$ from Example 2.13, it suffices to compute

$$\begin{aligned}
\mathsf{E}[X(X-1)] &= \sum_{n=0}^{\infty} n(n-1)\frac{\lambda^n e^{-\lambda}}{n!} \\
&= \sum_{n=2}^{\infty} \frac{\lambda^n e^{-\lambda}}{(n-2)!} \\
&= \lambda^2 e^{-\lambda} \sum_{n=2}^{\infty} \frac{\lambda^{n-2}}{(n-2)!}.
\end{aligned}$$

Making the change of summation $k = n - 2$, we have

$$\begin{aligned}
\mathsf{E}[X(X-1)] &= \lambda^2 e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= \lambda^2.
\end{aligned}$$

It follows that $\mathsf{E}[X^2] = \lambda^2 + \lambda$, and

$$\mathsf{var}(X) \;=\; \mathsf{E}[X^2] \Leftrightarrow (\mathsf{E}[X])^2 \;=\; (\lambda^2 + \lambda) \Leftrightarrow \lambda^2 \;=\; \lambda.$$

Thus, the Poisson(λ) random variable is unusual in that its mean and variance are the same.

Probability Generating Functions

Let X be a discrete random variable taking only nonnegative integer values. The **probability generating function** (pgf) of X is[4]

$$G_X(z) \;:=\; \mathsf{E}[z^X] \;=\; \sum_{n=0}^{\infty} z^n \wp(X = n).$$

Since for $|z| \leq 1$,

$$\begin{aligned} \left| \sum_{n=0}^{\infty} z^n \wp(X = n) \right| \;&\leq\; \sum_{n=0}^{\infty} |z^n \wp(X = n)| \\ &=\; \sum_{n=0}^{\infty} |z|^n \wp(X = n) \\ &\leq\; \sum_{n=0}^{\infty} \wp(X = n) \;=\; 1, \end{aligned}$$

G_X has a power series expansion with radius of convergence at least one. The reason that G_X is called the probability generating function is that the probabilities $\wp(X = k)$ can be obtained from G_X as follows. Writing

$$G_X(z) \;=\; \wp(X = 0) + z\,\wp(X = 1) + z^2\,\wp(X = 2) + \cdots,$$

we immediately see that $G_X(0) = \wp(X = 0)$. We also see that

$$G'_X(z) \;=\; \wp(X = 1) + 2z\,\wp(X = 2) + 3z^2\,\wp(X = 3) + \cdots.$$

Hence, $G'_X(0) = \wp(X = 1)$. Continuing in this way shows that

$$G_X^{(k)}(z)|_{z=0} \;=\; k!\,\wp(X = k),$$

or equivalently,

$$\frac{G_X^{(k)}(z)|_{z=0}}{k!} \;=\; \wp(X = k). \tag{2.5}$$

The probability generating function can also be used to find moments. Since

$$G'_X(z) \;=\; \sum_{n=1}^{\infty} n z^{n-1} \wp(X = n),$$

setting $z = 1$ yields

$$G'_X(z)|_{z=1} \;=\; \sum_{n=1}^{\infty} n\wp(X=n) \;=\; \mathsf{E}[X].$$

Similarly, since

$$G''_X(z) \;=\; \sum_{n=2}^{\infty} n(n\Leftrightarrow 1)z^{n-2}\wp(X=n),$$

setting $z = 1$ yields

$$G''_X(z)|_{z=1} \;=\; \sum_{n=2}^{\infty} n(n\Leftrightarrow 1)\wp(X=n) \;=\; \mathsf{E}[X(X\Leftrightarrow 1)] \;=\; \mathsf{E}[X^2]\Leftrightarrow \mathsf{E}[X].$$

In general, since

$$G_X^{(k)}(z) \;=\; \sum_{n=k}^{\infty} n(n\Leftrightarrow 1)\cdots(n\Leftrightarrow[k\Leftrightarrow 1])z^{n-k}\wp(X=n),$$

setting $z = 1$ yields

$$G_X^{(k)}(z)|_{z=1} \;=\; \mathsf{E}\big[X(X\Leftrightarrow 1)(X\Leftrightarrow 2)\cdots(X\Leftrightarrow[k\Leftrightarrow 1])\big].$$

The right-hand side is called the kth **factorial moment** of X.

***Example* 2.19.** Find the probability generating function of $X \sim \text{Poisson}(\lambda)$, and use the result to find $\mathsf{E}[X]$ and $\mathsf{var}(X)$.

Solution. Write

$$\begin{aligned} G_X(z) &= \mathsf{E}[z^X] \\ &= \sum_{n=0}^{\infty} z^n \wp(X=n) \\ &= \sum_{n=0}^{\infty} z^n \frac{\lambda^n e^{-\lambda}}{n!} \\ &= e^{-\lambda}\sum_{n=0}^{\infty} \frac{(z\lambda)^n}{n!} \\ &= e^{-\lambda}e^{z\lambda} \\ &= \exp[\lambda(z\Leftrightarrow 1)]. \end{aligned}$$

Now observe that $G'_X(z) = \exp[\lambda(z\Leftrightarrow 1)]\lambda$, and so $\mathsf{E}[X] = G'_X(1) = \lambda$. Since $G''_X(z) = \exp[\lambda(z\Leftrightarrow 1)]\lambda^2$, $\mathsf{E}[X^2]\Leftrightarrow\mathsf{E}[X] = \lambda^2$, and so $\mathsf{E}[X^2] = \lambda^2 + \lambda$. For the variance, write

$$\mathsf{var}(X) \;=\; \mathsf{E}[X^2]\Leftrightarrow(\mathsf{E}[X])^2 \;=\; (\lambda^2+\lambda)\Leftrightarrow\lambda^2 = \lambda.$$

Expectations of Products of Functions of Independent Random Variables

If X and Y are independent, then

$$\mathsf{E}[h(X)\,k(Y)] \;=\; \mathsf{E}[h(X)]\,\mathsf{E}[k(Y)]$$

for all functions $h(x)$ and $k(y)$. In other words, if X and Y are independent, then for any functions $h(x)$ and $k(y)$, the expectation of the product $h(X)k(Y)$ is equal to the product of the individual expectations. To derive this result, we use the fact that

$$\begin{aligned} p_{XY}(x_i, y_j) &:= \wp(X = x_i, Y = y_j) \\ &= \wp(X = x_i)\,\wp(Y = y_j), \quad \text{by independence,} \\ &= p_X(x_i)\,p_Y(y_j). \end{aligned}$$

Now write

$$\begin{aligned} \mathsf{E}[h(X)\,k(Y)] &= \sum_i \sum_j h(x_i)\,k(y_j)\,p_{XY}(x_i, y_j) \\ &= \sum_i \sum_j h(x_i)\,k(y_j)\,p_X(x_i)\,p_Y(y_j) \\ &= \sum_i h(x_i)\,p_X(x_i)\left[\sum_j k(y_j)\,p_Y(y_j)\right] \\ &= \mathsf{E}[h(X)]\,\mathsf{E}[k(Y)]. \end{aligned}$$

***Example* 2.20.** A certain communication network consists of n links. Suppose that each link goes down with probability p independently of the other links. For $k = 0, \ldots, n$, find the probability that exactly k out of the n links are down.

Solution. Let $X_i = 1$ if the ith link is down and $X_i = 0$ otherwise. Then the X_i are independent Bernoulli(p). Put $Y := X_1 + \cdots + X_n$. Then Y counts the number of links that are down. We first find the probability generating function of Y. Write

$$\begin{aligned} G_Y(z) &= \mathsf{E}[z^Y] \\ &= \mathsf{E}[z^{X_1 + \cdots + X_n}] \\ &= \mathsf{E}[z^{X_1} \cdots z^{X_n}] \\ &= \mathsf{E}[z^{X_1}] \cdots \mathsf{E}[z^{X_n}], \quad \text{by independence.} \end{aligned}$$

Now, the probability generating function of a Bernoulli(p) random variable is easily seen to be

$$\mathsf{E}[z^{X_i}] \;=\; z^0(1 \Leftrightarrow p) + z^1 p \;=\; (1 \Leftrightarrow p) + pz.$$

Thus,

$$G_Y(z) \;=\; [(1 \Leftrightarrow p) + pz]^n.$$

To apply (2.5), we need the derivatives of $G_Y(z)$. The first derivative is

$$G'_Y(z) \;=\; n[(1 \Leftrightarrow p) + pz]^{n-1} p,$$

and in general, the kth derivative is

$$G_Y^{(k)}(z) \;=\; n(n \Leftrightarrow 1) \cdots (n \Leftrightarrow [k \Leftrightarrow 1])\, [(1 \Leftrightarrow p) + pz]^{n-k} p^k.$$

It now follows that

$$\begin{aligned} \wp(Y = k) &= \frac{G_Y^{(k)}(0)}{k!} \\ &= \frac{n(n \Leftrightarrow 1) \cdots (n \Leftrightarrow [k \Leftrightarrow 1])}{k!} (1 \Leftrightarrow p)^{n-k} p^k \\ &= \frac{n!}{k!(n \Leftrightarrow k)!} p^k (1 \Leftrightarrow p)^{n-k}. \end{aligned}$$

Since the formula for $G_Y(z)$ is a polynomial of degree n, $G_Y^{(k)}(z) = 0$ for all $k > n$. Thus, $\wp(Y = k) = 0$ for $k > n$.

The preceding random variable Y is an example of a **binomial**(n, p) random variable. (An alternative derivation using techniques from Section 2.4 is given in the Notes.[5]) The binomial random variable counts how many times an event has occurred. Its probability mass function is usually written using the notation

$$p_Y(k) \;=\; \binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k}, \quad k = 0, \ldots, n,$$

where the symbol $\binom{n}{k}$ is read "n choose k," and is defined by

$$\binom{n}{k} \;:=\; \frac{n!}{k!(n \Leftrightarrow k)!}.$$

In MATLAB, $\binom{n}{k} =$ `nchoosek(n,k)`.

For any discrete random variable Y taking only nonnegative integer values, we always have

$$G_Y(1) \;=\; \left(\sum_{k=0}^{\infty} z^k \, p_Y(k) \right) \Bigg|_{z=1} \;=\; \sum_{k=0}^{\infty} p_Y(k) \;=\; 1.$$

For the binomial random variable Y this says

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k} \;=\; 1,$$

when p is any number satisfying $0 \le p \le 1$. More generally, suppose a and b are any nonnegative numbers with $a + b > 0$. Put $p = a/(a+b)$. Then $0 \le p \le 1$, and $1 \Leftrightarrow p = b/(a+b)$. It follows that

$$\sum_{k=0}^{n} \binom{n}{k} \left(\frac{a}{a+b}\right)^k \left(\frac{b}{a+b}\right)^{n-k} \;=\; 1.$$

Multiplying both sides by $(a+b)^k(a+b)^{n-k} = (a+b)^n$ yields the **binomial theorem**,

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \;=\; (a+b)^n.$$

Remark. The above derivation is for nonnegative a and b. The binomial theorem actually holds for complex a and b. This is easy to derive using induction on n along with the easily verified identity

$$\binom{n}{k \Leftarrow 1} + \binom{n}{k} \;=\; \binom{n+1}{k}.$$

On account of the binomial theorem, the quantity $\binom{n}{k}$ is sometimes called the **binomial coefficient**. We also point out that the binomial coefficients can be read off from the nth row of **Pascal's triangle** in Figure 2.4. Noting that the top row is row 0, it is immediate, for example, that

$$(a+b)^5 \;=\; a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5.$$

To generate the triangle, observe that except for the entries that are ones, each entry is equal to the sum of the two numbers above it to the left and right.

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

Figure 2.4. Pascal's triangle.

Binomial Random Variables and Combinations

We showed in Example 2.20 that if $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), then $Y := X_1 + \cdots + X_n$ is binomial(n, p). We can use this result to show that the

binomial coefficient $\binom{n}{k}$ is equal to the number of n-bit words with exactly k ones and $n \Leftrightarrow k$ zeros. On the one hand, $\wp(Y = k) = \binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k}$. On the other hand, observe that $Y = k$ if and only if

$$X_1 = i_1, \ldots, X_n = i_n$$

for some n-tuple $(i_1, \ldots, i_n)$ of ones and zeros with exactly k ones and $n \Leftrightarrow k$ zeros. Denoting this set of n-tuples by $B_{n,k}$, we can write

$$\begin{aligned} \wp(Y = k) &= \wp\Bigg(\bigcup_{(i_1,\ldots,i_n) \in B_{n,k}} \{X_1 = i_1, \ldots, X_n = i_n\} \Bigg) \\ &= \sum_{(i_1,\ldots,i_n) \in B_{n,k}} \wp(X_1 = i_1, \ldots, X_n = i_n) \\ &= \sum_{(i_1,\ldots,i_n) \in B_{n,k}} \prod_{\nu=1}^{n} \wp(X_\nu = i_\nu). \end{aligned}$$

Now, each factor is p if $i_\nu = 1$ and $1 \Leftrightarrow p$ if $i_\nu = 0$. Since each n-tuple belongs to $B_{n,k}$, we must have k factors being p and $n \Leftrightarrow k$ factors being $1 \Leftrightarrow p$. Thus,

$$\wp(Y = k) = \sum_{(i_1,\ldots,i_n) \in B_{n,k}} p^k (1 \Leftrightarrow p)^{n-k}.$$

Since all terms in the above sum are the same,

$$\wp(Y = k) = |B_{n,k}|\, p^k (1 \Leftrightarrow p)^{n-k}.$$

It follows that $|B_{n,k}| = \binom{n}{k}$.

Consider n objects $O_1, \ldots, O_n$. How many ways can we select exactly k of the n objects if the order of selection is not important? Each such selection is

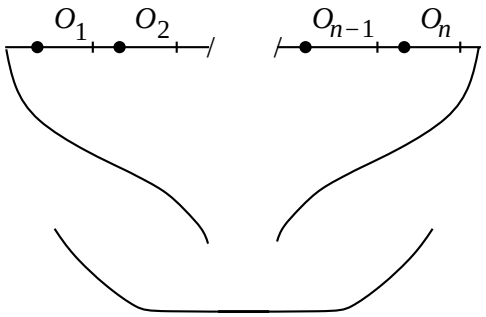

Figure 2.5. There are n objects arranged over trap doors. If k doors are opened (and $n - k$ shut), this device allows the corresponding combination of k of the n objects to fall into the bowl below.

called a **combination**. One way to answer this question is to think of arranging the n objects over trap doors as shown in Figure 2.5. If k doors are opened (and $n \Leftrightarrow k$ shut), this device allows the corresponding combination of k of the n objects to fall into the bowl below. The number of ways to select the open and shut doors is exactly the number of n-bit words with k ones and $n \Leftrightarrow k$ zeros. As argued above, this number is $\binom{n}{k}$.

***Example* 2.21.** A 12-person jury is to be selected from a group of 20 potential jurors. How many different juries are possible?

Solution. There are

$$\binom{20}{12} = \frac{20!}{12!\,8!} = 125970$$

different juries.

***Example* 2.22.** A 12-person jury is to be selected from a group of 20 potential jurors of which 11 are men and 9 are women. How many 12-person juries are there with 5 men and 7 women?

Solution. There are $\binom{11}{5}$ ways to choose the 5 men, and there are $\binom{9}{7}$ ways to choose the 7 women. Hence, there are

$$\binom{11}{5}\binom{9}{7} = \frac{11!}{5!\,6!} \cdot \frac{9!}{7!\,2!} = 16632$$

possible juries with 5 men and 7 women.

***Example* 2.23.** An urn contains 11 green balls and 9 red balls. If 12 balls are chosen at random, what is the probability of choosing exactly 5 green balls and 7 red balls?

Solution. Since there are $\binom{20}{12}$ ways to choose any 12 balls, we envision a sample space Ω with $|\Omega| = \binom{20}{12}$. We are interested in a subset of Ω, call it $B_{\mathrm{g,r}}(5,7)$, corresponding to those selections of 12 balls such that 5 are green and 7 are red. Thus,

$$|B_{\mathrm{g,r}}(5,7)| = \binom{11}{5}\binom{9}{7}.$$

Since all selections of 12 balls are equally likely, the probability we seek is

$$\frac{|B_{\mathrm{g,r}}(5,7)|}{|\Omega|} = \frac{\binom{11}{5}\binom{9}{7}}{\binom{20}{12}} = \frac{16632}{125970} \approx 0.132.$$

Poisson Approximation of Binomial Probabilities

If we let $\lambda := np$, then the probability generating function of a binomial(n,p) random variable can be written as

$$\begin{aligned}[(1 \Leftrightarrow p) + pz]^n &= [1 + p(z \Leftrightarrow 1)]^n \\ &= \left[1 + \frac{\lambda(z \Leftrightarrow 1)}{n}\right]^n.\end{aligned}$$

From calculus, recall the formula

$$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x.$$

So, for large n,

$$\left[1 + \frac{\lambda(z \Leftrightarrow 1)}{n}\right]^n \approx \exp[\lambda(z \Leftrightarrow 1)],$$

which is the probability generating function of a Poisson(λ) random variable (Example 2.19). In making this approximation, n should be large compared to $\lambda(z \Leftrightarrow 1)$. Since $\lambda := np$, as n becomes large, so does $\lambda(z \Leftrightarrow 1)$. To keep the size of λ small enough to be useful, we should keep p small. Under this assumption, the binomial(n,p) probability generating function is close to the Poisson(np) probability generating function. This suggests the approximation[‡]

$$\binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k} \approx \frac{(np)^k e^{-np}}{k!}, \quad n \text{ large, } p \text{ small.}$$

2.3. The Weak Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of random variables with a common mean $\mathsf{E}[X_i] = m$ for all i. In practice, since we do not know m, we use the numerical average, or **sample mean**,

$$M_n := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

in place of the true, but unknown value, m. Can this procedure of using M_n as an estimate of m be justified in some sense?

***Example* 2.24.** You are given a coin which may or may not be fair, and you want to determine the probability of heads, p. If you toss the coin n times and use the fraction of times that heads appears as an estimate of p, how does this fit into the above framework?

Solution. Let $X_i = 1$ if the ith toss results in heads, and let $X_i = 0$ otherwise. Then $\wp(X_i = 1) = p$ and $m := \mathsf{E}[X_i] = p$ as well. Note that $X_1 + \cdots + X_n$ is the number of heads, and M_n is the fraction of heads. Are we justified in using M_n as an estimate of p?

[‡]This approximation is justified rigorously in Problems 15 and 16(a) in Chapter 11. It is also derived directly without probability generating functions in Problem 17 in Chapter 11.

One way to answer these questions is with a **weak law of large numbers (WLLN)**. A weak law of large numbers gives conditions under which

$$\lim_{n\to\infty} \wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;=\; 0$$

for every $\varepsilon > 0$. This is a complicated formula. However, it can be interpreted as follows. Suppose that based on physical considerations, m is between 30 and 70. Let us agree that if M_n is within $\varepsilon = 1/2$ of m, we are "close enough" to the unknown value m. For example, if $M_n = 45.7$, and if we know that M_n is within $1/2$ of m, then m is between 45.2 and 46.2. Knowing this would be an improvement over the starting point $30 \leq m \leq 70$. So, if $|M_n \Leftrightarrow m| < \varepsilon$, we are "close enough," while if $|M_n \Leftrightarrow m| \geq \varepsilon$ we are not "close enough." A weak law says that by making n large (averaging lots of measurements), the probability of not being close enough can be made as small as we like; equivalently, the probability of being close enough can be made as close to one as we like. For example, if $\wp(|M_n \Leftrightarrow m| \geq \varepsilon) \leq 0.1$, then

$$\wp(|M_n \Leftrightarrow m| < \varepsilon) \;=\; 1 \Leftrightarrow \wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;\geq\; 0.9,$$

and we would be 90% sure that M_n is "close enough" to the true, but unknown, value of m.

Before giving conditions under which the weak law holds, we need to introduce the notion of uncorrelated random variables. Also, in order to prove the weak law, we need to first derive Markov's inequality and Chebyshev's inequality.

Uncorrelated Random Variables

Before giving conditions under which the weak law holds, we need to introduce the notion of **uncorrelated random variables**. A pair of random variables, say X and Y, is said to be uncorrelated if

$$\mathsf{E}[XY] \;=\; \mathsf{E}[X]\,\mathsf{E}[Y].$$

If X and Y are independent, then they are uncorrelated. Intuitively, the property of being uncorrelated is weaker than the property of independence. Recall that for independent random variables,

$$\mathsf{E}[h(X)\,k(Y)] \;=\; \mathsf{E}[h(X)]\,\mathsf{E}[k(Y)]$$

for *all* functions $h(x)$ and $k(y)$, while for uncorrelated random variables, we only require that that this hold for $h(x) = x$ and $k(y) = y$. For an example of uncorrelated random variables that are not independent, see Problem 41.

***Example* 2.25.** If $m_X := \mathsf{E}[X]$ and $m_Y := \mathsf{E}[Y]$, show that X and Y are uncorrelated if and only if

$$\mathsf{E}[(X \Leftrightarrow m_X)(Y \Leftrightarrow m_Y)] \;=\; 0.$$

Solution. The result is obvious if we expand

$$\begin{aligned}\mathsf{E}[(X \Leftrightarrow m_X)(Y \Leftrightarrow m_Y)] &= \mathsf{E}[XY \Leftrightarrow m_X Y \Leftrightarrow X m_Y + m_X m_Y] \\ &= \mathsf{E}[XY] \Leftrightarrow m_X \mathsf{E}[Y] \Leftrightarrow \mathsf{E}[X] m_Y + m_X m_Y \\ &= \mathsf{E}[XY] \Leftrightarrow m_X m_Y \Leftrightarrow m_X m_Y + m_X m_Y \\ &= \mathsf{E}[XY] \Leftrightarrow m_X m_Y.\end{aligned}$$

***Example* 2.26.** Let $X_1, X_2, \ldots$ be a sequence of uncorrelated random variables; more precisely, for $i \neq j$, X_i and X_j are uncorrelated. Show that

$$\mathsf{var}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} \mathsf{var}(X_i).$$

In other words, for uncorrelated random variables, the variance of the sum is the sum of the variances.

Solution. Let $m_i := \mathsf{E}[X_i]$ and $m_j := \mathsf{E}[X_j]$. Then uncorrelated means that

$$\mathsf{E}[(X_i \Leftrightarrow m_i)(X_j \Leftrightarrow m_j)] \;=\; 0 \quad \text{for all } i \neq j.$$

Put

$$X \;:=\; \sum_{i=1}^{n} X_i.$$

Then

$$\mathsf{E}[X] \;=\; \mathsf{E}\left[\sum_{i=1}^{n} X_i\right] \;=\; \sum_{i=1}^{n} \mathsf{E}[X_i] \;=\; \sum_{i=1}^{n} m_i,$$

and

$$X \Leftrightarrow \mathsf{E}[X] \;=\; \sum_{i=1}^{n} X_i \Leftrightarrow \sum_{i=1}^{n} m_i \;=\; \sum_{i=1}^{n} (X_i \Leftrightarrow m_i).$$

Now write

$$\begin{aligned}\mathsf{var}(X) &= \mathsf{E}[(X \Leftrightarrow \mathsf{E}[X])^2] \\ &= \mathsf{E}\left[\left(\sum_{i=1}^{n} (X_i \Leftrightarrow m_i)\right)\left(\sum_{j=1}^{n} (X_j \Leftrightarrow m_j)\right)\right] \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} \mathsf{E}[(X_i \Leftrightarrow m_i)(X_j \Leftrightarrow m_j)]\right).\end{aligned}$$

For fixed i, consider the sum over j. There is only one term in this sum for which $j \neq i$, and that is the term $j = i$. All the other terms are zero because

X_i and X_j are uncorrelated. Hence,

$$\begin{aligned} \mathsf{var}(X) &= \sum_{i=1}^{n} \Big(\mathsf{E}[(X_i \Leftrightarrow m_i)(X_i \Leftrightarrow m_i)] \Big) \\ &= \sum_{i=1}^{n} \mathsf{E}[(X_i \Leftrightarrow m_i)^2] \\ &= \sum_{i=1}^{n} \mathsf{var}(X_i). \end{aligned}$$

Markov's Inequality

If X is a nonnegative random variable, we show that for any $a > 0$,

$$\wp(X \geq a) \;\leq\; \frac{\mathsf{E}[X]}{a}.$$

This relation is known as **Markov's inequality**. Since we always have $\wp(X \geq a) \leq 1$, Markov's inequality is useful only when the right-hand side is less than one.

Consider the random variable $a I_{[a,\infty)}(X)$. This random variable takes the value zero if $X < a$, and it takes the value a if $X \geq a$. Hence,

$$\mathsf{E}[a I_{[a,\infty)}(X)] \;=\; 0 \cdot \wp(X < a) + a \cdot \wp(X \geq a) \;=\; a\,\wp(X \geq a).$$

Now observe that

$$a I_{[a,\infty)}(X) \;\leq\; X.$$

To see this, note that the left-hand side is either zero or a. If it is zero, there is no problem because X is a nonnegative random variable. If the left-hand side is a, then we must have $I_{[a,\infty)}(X) = 1$; but this means that $X \geq a$. Next take expectations of both sides to obtain

$$a\,\wp(X \geq a) \;\leq\; \mathsf{E}[X].$$

Dividing both sides by a yields Markov's inequality.

Chebyshev's Inequality

For any random variable Y and any $a > 0$, we show that

$$\wp(|Y| \geq a) \;\leq\; \frac{\mathsf{E}[Y^2]}{a^2}.$$

This result, known as **Chebyshev's inequality**, is an easy consequence of Markov's inequality. As in the case of Markov's inequality, it is useful only when the right-hand side is less than one.

To prove Chebyshev's inequality, just note that

$$\{|Y| \geq a\} \;=\; \{|Y|^2 \geq a^2\}.$$

Since the above two events are equal, they have the same probability. Hence,

$$\wp(|Y| \geq a) \;=\; \wp(|Y|^2 \geq a^2) \;\leq\; \frac{\mathsf{E}[|Y|^2]}{a^2},$$

where the last step follows by Markov's inequality.

The following special cases of Chebyshev's inequality are sometimes of interest. If $m := \mathsf{E}[X]$ is finite, then taking $Y = |X \Leftrightarrow m|$ yields

$$\wp(|X \Leftrightarrow m| \geq a) \;\leq\; \frac{\mathsf{var}(X)}{a^2}.$$

If $\sigma^2 := \mathsf{var}(X)$ is also finite, taking $a = k\sigma$ yields

$$\wp(|X \Leftrightarrow m| \geq k\sigma) \;\leq\; \frac{1}{k^2}.$$

Conditions for the Weak Law

We now give sufficient conditions for a version of the **weak law of large numbers (WLLN)**. Suppose that the X_i all have the same mean m and the same variance σ^2. Assume also that the X_i are uncorrelated random variables. Then for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;=\; 0.$$

This is an immediate consequence of the following two facts. First, by Chebyshev's inequality,

$$\wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;\leq\; \frac{\mathsf{var}(M_n)}{\varepsilon^2}.$$

Second,

$$\mathsf{var}(M_n) \;=\; \mathsf{var}\left(\frac{1}{n}\sum_{n=1}^{n} X_i\right) \;=\; \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathsf{var}(X_i) \;=\; \frac{n\sigma^2}{n^2} \;=\; \frac{\sigma^2}{n}. \qquad (2.6)$$

Thus,

$$\wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;\leq\; \frac{\sigma^2}{n\varepsilon^2}, \qquad (2.7)$$

which goes to zero as $n \to \infty$. Note that the bound $\sigma^2/n\varepsilon^2$ can be used to select a suitable value of n.

***Example* 2.27.** Given ε and σ^2, determine how large n should be so that the probability that M_n is within ε of m is at least 0.9.

Solution. We want to have

$$\wp(|M_n \Leftrightarrow m| < \varepsilon) \;\geq\; 0.9.$$

Rewrite this as

$$1 \Leftrightarrow \wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;\geq\; 0.9,$$

or

$$\wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;\leq\; 0.1.$$

By (2.7), it suffices to take

$$\frac{\sigma^2}{n\varepsilon^2} \;\leq\; 0.1,$$

or $n \geq 10\sigma^2/\varepsilon^2$.

2.4. Conditional Probability

For conditional probabilities involving random variables, we use the notation

$$\begin{aligned}
\wp(X \in B | Y \in C) &:= \wp(\{X \in B\} | \{Y \in C\}) \\
&= \frac{\wp(\{X \in B\} \cap \{Y \in C\})}{\wp(\{Y \in C\})} \\
&= \frac{\wp(X \in B, Y \in C)}{\wp(Y \in C)}.
\end{aligned}$$

For discrete random variables, we define the **conditional probability mass functions**,

$$\begin{aligned}
p_{X|Y}(x_i|y_j) &:= \wp(X = x_i | Y = y_j) \\
&= \frac{\wp(X = x_i, Y = y_j)}{\wp(Y = y_j)} \\
&= \frac{p_{XY}(x_i, y_j)}{p_Y(y_j)},
\end{aligned}$$

and

$$\begin{aligned}
p_{Y|X}(y_j|x_i) &:= \wp(Y = y_j | X = x_i) \\
&= \frac{\wp(X = x_i, Y = y_j)}{\wp(X = x_i)} \\
&= \frac{p_{XY}(x_i, y_j)}{p_X(x_i)}.
\end{aligned}$$

We call $p_{X|Y}$ the conditional probability mass function of X given Y. Similarly, $p_{Y|X}$ is called the conditional pmf of Y given X.

Conditional pmfs are important because we can use them to compute conditional probabilities just as we use marginal pmfs to compute ordinary probabilities. For example,

$$\wp(Y \in C|X = x_k) \;=\; \sum_j I_C(y_j) p_{Y|X}(y_j|x_k).$$

This formula is derived by taking $B = \{x_k\}$ in (2.3), and then dividing the result by $\wp(X = x_k) = p_X(x_k)$.

***Example* 2.28.** To transmit message i using an optical communication system, light of intensity λ_i is directed at a photodetector. When light of intensity λ_i strikes the photodetector, the number of photoelectrons generated is a Poisson(λ_i) random variable. Find the conditional probability that the number of photoelectrons observed at the photodetector is less than 2 given that message i was sent.

Solution. Let X denote the message to be sent, and let Y denote the number of photoelectrons generated by the photodetector. The problem statement is telling us that

$$\wp(Y = n|X = i) \;=\; \frac{\lambda_i^n e^{-\lambda_i}}{n!}, \quad n = 0, 1, 2, \ldots.$$

The conditional probability to be calculated is

$$\begin{aligned} \wp(Y < 2|X = i) &= \wp(Y = 0 \text{ or } Y = 1|X = i) \\ &= \wp(Y = 0|X = i) + \wp(Y = 1|X = i) \\ &= e^{-\lambda_i} + \lambda_i e^{-\lambda_i}. \end{aligned}$$

***Example* 2.29.** For the random variables X and Y used in the solution of the previous example, write down their joint pmf if $X \sim \text{geometric}_0(p)$.

Solution. The joint pmf is

$$p_{XY}(i, n) \;=\; p_X(i)\, p_{Y|X}(n|i) \;=\; (1 \Leftrightarrow p) p^i \, \frac{\lambda_i^n e^{-\lambda_i}}{n!},$$

for $i, n \geq 0$, and $p_{XY}(i, n) = 0$ otherwise.

The Law of Total Probability

Let $A \subset \Omega$ be any event, and let X be any discrete random variable taking distinct values x_i. Then the events

$$B_i \;:=\; \{X = x_i\} \;=\; \{\omega \in \Omega : X(\omega) = x_i\}.$$

are pairwise disjoint, and $\sum_i \wp(B_i) = \sum_i \wp(X = x_i) = 1$. The law of total probability as in (1.13) yields

$$\wp(A) = \sum_i \wp(A \cap B_i) = \sum_i \wp(A|X = x_i)\wp(X = x_i). \tag{2.8}$$

Hence, we call (2.8) the law of total probability as well.

Now suppose that Y is an arbitrary random variable. Taking $A = \{Y \in C\}$, where $C \subset \mathbb{R}$, yields

$$\wp(Y \in C) = \sum_i \wp(Y \in C|X = x_i)\wp(X = x_i).$$

If Y is a discrete random variable taking distinct values y_j, then setting $C = \{y_j\}$ yields

$$\begin{aligned} \wp(Y = y_j) &= \sum_i \wp(Y = y_j|X = x_i)\wp(X = x_i) \\ &= \sum_i p_{Y|X}(y_j|x_i)p_X(x_i). \end{aligned}$$

***Example* 2.30.** Radioactive samples give off alpha-particles at a rate based on the size of the sample. For a sample of size k, suppose that the number of particles observed is a Poisson random variable Y with parameter k. If the sample size is a geometric$_1(p)$ random variable X, find $\wp(Y = 0)$ and $\wp(X = 1|Y = 0)$.

Solution. The first step is to realize that the problem statement is telling us that as a function of n, $\wp(Y = n|X = k)$ is the Poisson pmf with parameter k. In other words,

$$\wp(Y = n|X = k) = \frac{k^n e^{-k}}{n!}, \quad n = 0, 1, \ldots.$$

In particular, note that $\wp(Y = 0|X = k) = e^{-k}$. Now use the law of total probability to write

$$\begin{aligned} \wp(Y = 0) &= \sum_{k=1}^{\infty} \wp(Y = 0|X = k) \cdot \wp(X = k) \\ &= \sum_{k=1}^{\infty} e^{-k} \cdot (1 \Leftrightarrow p)p^{k-1} \\ &= \frac{1 \Leftrightarrow p}{p} \sum_{k=1}^{\infty} (p/e)^k \\ &= \frac{1 \Leftrightarrow p}{p} \frac{p/e}{1 \Leftrightarrow p/e} = \frac{1 \Leftrightarrow p}{e \Leftrightarrow p}. \end{aligned}$$

Next,

$$\begin{aligned}
\wp(X=1|Y=0) &= \frac{\wp(X=1,Y=0)}{\wp(Y=0)} \\
&= \frac{\wp(Y=0|X=1)\wp(X=1)}{\wp(Y=0)} \\
&= e^{-1}\cdot(1 \Leftrightarrow p)\cdot\frac{e \Leftrightarrow p}{1 \Leftrightarrow p} \\
&= 1 \Leftrightarrow p/e.
\end{aligned}$$

***Example* 2.31.** A certain electric eye employs a photodetector whose efficiency occasionally drops in half. When operating properly, the detector outputs photoelectrons according to a Poisson(λ) pmf. When the detector malfunctions, it outputs photoelectrons according to a Poisson($\lambda/2$) pmf. Let $p<1$ denote the probability that the detector is operating properly. Find the pmf of the observed number of photoelectrons. Also find the conditional probability that the circuit is malfunctioning given that n output photoelectrons are observed.

Solution. Let Y denote the detector output, and let $X=1$ indicate that the detector is operating properly. Let $X=0$ indicate that it is malfunctioning. Then the problem statement is telling us that $\wp(X=1)=p$ and

$$\wp(Y=n|X=1) = \frac{\lambda^n e^{-\lambda}}{n!} \quad \text{and} \quad \wp(Y=n|X=0) = \frac{(\lambda/2)^n e^{-\lambda/2}}{n!}.$$

Now, using the law of total probability,

$$\begin{aligned}
\wp(Y=n) &= \wp(Y=n|X=1)\wp(X=1)+\wp(Y=n|X=0)\wp(X=0) \\
&= \frac{\lambda^n e^{-\lambda}}{n!}p + \frac{(\lambda/2)^n e^{-\lambda/2}}{n!}(1 \Leftrightarrow p).
\end{aligned}$$

This is the pmf of the observed number of photoelectrons.

The above formulas can be used to find $\wp(X=0|Y=n)$. Write

$$\begin{aligned}
\wp(X=0|Y=n) &= \frac{\wp(X=0,Y=n)}{\wp(Y=n)} \\
&= \frac{\wp(Y=n|X=0)\wp(X=0)}{\wp(Y=n)} \\
&= \frac{\dfrac{(\lambda/2)^n e^{-\lambda/2}}{n!}(1 \Leftrightarrow p)}{\dfrac{\lambda^n e^{-\lambda}}{n!}p + \dfrac{(\lambda/2)^n e^{-\lambda/2}}{n!}(1 \Leftrightarrow p)} \\
&= \frac{1}{\dfrac{2^n e^{-\lambda/2}p}{(1 \Leftrightarrow p)}+1},
\end{aligned}$$

which is clearly a number between zero and one as a probability should be. Notice that as we observe a greater output $Y = n$, the conditional probability that the detector is malfunctioning decreases.

The Substitution Law

It is often the case that Z is a function of X and some other discrete random variable Y, say $Z = X + Y$, and we are interested in $\wp(Z = z)$. In this case, the law of total probability becomes

$$\begin{aligned}\wp(Z = z) &= \sum_i \wp(Z = z | X = x_i)\wp(X = x_i). \\ &= \sum_i \wp(X + Y = z | X = x_i)\wp(X = x_i).\end{aligned}$$

We now claim that

$$\wp(X + Y = z | X = x_i) = \wp(x_i + Y = z | X = x_i).$$

This property is known as the **substitution law** of conditional probability. To derive it, we need the observation

$$\{X + Y = z\} \cap \{X = x_i\} = \{x_i + Y = z\} \cap \{X = x_i\}.$$

From this we see that

$$\begin{aligned}\wp(X + Y = z | X = x_i) &= \frac{\wp(\{X + Y = z\} \cap \{X = x_i\})}{\wp(\{X = x_i\})} \\ &= \frac{\wp(\{x_i + Y = z\} \cap \{X = x_i\})}{\wp(\{X = x_i\})} \\ &= \wp(x_i + Y = z | X = x_i).\end{aligned}$$

Writing

$$\wp(x_i + Y = z | X = x_i) = \wp(Y = z \Leftrightarrow x_i | X = x_i),$$

we can make further simplifications if X and Y are independent. In this case,

$$\begin{aligned}\wp(Y = z \Leftrightarrow x_i | X = x_i) &= \frac{\wp(Y = z \Leftrightarrow x_i, X = x_i)}{\wp(X = x_i)} \\ &= \frac{\wp(Y = z \Leftrightarrow x_i)\wp(X = x_i)}{\wp(X = x_i)} \\ &= \wp(Y = z \Leftrightarrow x_i).\end{aligned}$$

Thus, when X and Y are independent, we can write

$$\wp(Y = z \Leftrightarrow x_i | X = x_i) = \wp(Y = z \Leftrightarrow x_i),$$

and we say that we "drop the conditioning."

***Example* 2.32** (Signal in Additive Noise). A random, integer-valued signal X is transmitted over a channel subject to independent, additive, integer-valued noise Y. The received signal is $Z = X + Y$. To estimate X based on the received value Z, the system designer wants to use the conditional pmf $p_{X|Z}$. Our problem is to find this conditional pmf.

Solution. Let X and Y be independent, discrete, integer-valued random variables with pmfs p_X and p_Y, respectively. Put $Z := X + Y$. We begin by writing out the formula for the desired pmf

$$\begin{aligned} p_{X|Z}(i|j) &= \wp(X = i|Z = j) \\ &= \frac{\wp(X = i, Z = j)}{\wp(Z = j)} \\ &= \frac{\wp(Z = j|X = i)\wp(X = i)}{\wp(Z = j)} \\ &= \frac{\wp(Z = j|X = i)p_X(i)}{\wp(Z = j)}. \end{aligned} \tag{2.9}$$

To continue the analysis, we use the substitution law followed by independence to write

$$\begin{aligned} \wp(Z = j|X = i) &= \wp(X + Y = j|X = i) \\ &= \wp(i + Y = j|X = i) \\ &= \wp(Y = j \Leftrightarrow i|X = i) \\ &= \wp(Y = j \Leftrightarrow i) \\ &= p_Y(j \Leftrightarrow i). \end{aligned} \tag{2.10}$$

This result can also be combined with the law of total probability to compute the denominator in (2.9). Just write

$$p_Z(j) = \sum_i \wp(Z = j|X = i)\wp(X = i) = \sum_i p_Y(j \Leftrightarrow i)p_X(i). \tag{2.11}$$

In other words, if X and Y are independent, discrete, integer-valued random variables, the pmf of $Z = X + Y$ is the discrete **convolution** of p_X and p_Y.

It now follows that

$$p_{X|Z}(i|j) = \frac{p_Y(j \Leftrightarrow i)p_X(i)}{\sum_k p_Y(j \Leftrightarrow k)p_X(k)},$$

where in the denominator we have changed the dummy index of summation to k to avoid confusion with the i in the numerator.

The Poisson(λ) random variable is a good model for the number of photoelectrons generated in a photodetector when the incident light intensity is λ.

Now suppose that an additional light source of intensity μ is also directed at the photodetector. Then we expect that the number of photoelectrons generated should be related to the total light intensity $\lambda+\mu$. The next example illustrates the corresponding probabilistic model.

***Example* 2.33.** If X and Y are independent Poisson random variables with respective parameters λ and μ, use the results of the preceding example to show that $Z := X + Y$ is Poisson$(\lambda+\mu)$. Also show that as a function of i, $p_{X|Z}(i|j)$ is a binomial$(j, \lambda/(\lambda+\mu))$ pmf.

Solution. To find $p_Z(j)$, we apply (2.11) as follows. Since $p_X(i) = 0$ for $i < 0$ and since $p_Y(j \ominus i) = 0$ for $j < i$, (2.11) becomes

$$\begin{aligned} p_Z(j) &= \sum_{i=0}^{j} \frac{\lambda^i e^{-\lambda}}{i!} \cdot \frac{\mu^{j-i} e^{-\mu}}{(j \ominus i)!} \\ &= \frac{e^{-(\lambda+\mu)}}{j!} \sum_{i=0}^{j} \frac{j!}{i!(j \ominus i)!} \lambda^i \mu^{j-i} \\ &= \frac{e^{-(\lambda+\mu)}}{j!} \sum_{i=0}^{j} \binom{j}{i} \lambda^i \mu^{j-i} \\ &= \frac{e^{-(\lambda+\mu)}}{j!} (\lambda+\mu)^j, \quad j = 0, 1, \ldots, \end{aligned}$$

where the last step follows by the binomial theorem.

Our second task is to compute

$$\begin{aligned} p_{X|Z}(i|j) &= \frac{\wp(Z=j|X=i)\wp(X=i)}{\wp(Z=j)} \\ &= \frac{\wp(Z=j|X=i)p_X(i)}{p_Z(j)}. \end{aligned}$$

Since we have already found $p_Z(j)$, all we need is $\wp(Z=j|X=i)$, which, using (2.10), is simply $p_Y(j \ominus i)$. Thus,

$$\begin{aligned} p_{X|Z}(i|j) &= \frac{\mu^{j-i} e^{-\mu}}{(j \ominus i)!} \cdot \frac{\lambda^i e^{-\lambda}}{i!} \Bigg/ \left[\frac{e^{-(\lambda+\mu)}}{j!} (\lambda+\mu)^j \right] \\ &= \frac{\lambda^i \mu^{j-i}}{(\lambda+\mu)^j} \binom{j}{i} \\ &= \binom{j}{i} \left[\frac{\lambda}{(\lambda+\mu)} \right]^i \left[\frac{\mu}{(\lambda+\mu)} \right]^{j-i}, \end{aligned}$$

for $i = 0, \ldots, j$.

2.5. Conditional Expectation

Just as we developed expectation for discrete random variables in Section 2.2, including the law of the unconscious statistician, we can develop conditional expectation in the same way. This leads to the formula

$$\mathsf{E}[g(Y)|X = x_i] \;=\; \sum_j g(y_j)\, p_{Y|X}(y_j|x_i). \tag{2.12}$$

***Example* 2.34.** The random number Y of alpha-particles given off by a radioactive sample is Poisson(k) given that the sample size $X = k$. Find $\mathsf{E}[Y|X = k]$.

Solution. We must compute

$$\mathsf{E}[Y|X = k] \;=\; \sum_n n\, \wp(Y = n|X = k),$$

where (cf. Example 2.30)

$$\wp(Y = n|X = k) \;=\; \frac{k^n e^{-k}}{n!}, \quad n = 0, 1, \ldots.$$

Hence,

$$\mathsf{E}[Y|X = k] \;=\; \sum_{n=0}^{\infty} n \frac{k^n e^{-k}}{n!}.$$

Now observe that the right-hand side is exactly ordinary expectation of a Poisson random variable with parameter k (cf. the calculation in Example 2.13). Therefore, $\mathsf{E}[Y|X = k] = k$.

***Example* 2.35.** Suppose that given $Y = n$, $X \sim$ binomial(n, p). Find $\mathsf{E}[X|Y = n]$.

Solution. We must compute

$$\mathsf{E}[X|Y = n] \;=\; \sum_{k=0}^{n} k \wp(X = k|Y = n),$$

where

$$\wp(X = k|Y = n) \;=\; \binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k}.$$

Hence,

$$\mathsf{E}[X|Y = n] \;=\; \sum_{k=0}^{n} k \binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k}.$$

Now observe that the right-hand side is exactly the ordinary expectation of a binomial(n, p) random variable. It is shown in Problem 20 that the mean of such a random variable is np. Therefore, $\mathsf{E}[X|Y = n] = np$.

Substitution Law for Conditional Expectation

For functions of two variables, we have the following conditional law of the unconscious statistician,

$$\mathsf{E}[g(X,Y)|X=x_i] \;=\; \sum_k \sum_j g(x_k,y_j)\, p_{XY|X}(x_k,y_j|x_i).$$

However,

$$\begin{aligned} p_{XY|X}(x_k,y_j|x_i) &= \wp(X=x_k, Y=y_j|X=x_i) \\ &= \frac{\wp(X=x_k, Y=y_j, X=x_i)}{\wp(X=x_i)}. \end{aligned}$$

Now, when $k \neq i$, the intersection

$$\{X=x_k\} \cap \{Y=y_j\} \cap \{X=x_i\}$$

is empty, and has zero probability. Hence, the numerator above is zero for $k \neq i$. When $k=i$, the above intersections reduce to $\{X=x_i\} \cap \{Y=y_j\}$, and so

$$p_{XY|X}(x_k,y_j|x_i) \;=\; p_{Y|X}(y_j|x_i), \quad \text{for } k=i.$$

It now follows that

$$\begin{aligned} \mathsf{E}[g(X,Y)|X=x_i] &= \sum_j g(x_i,y_j)\, p_{Y|X}(y_j|x_i) \\ &= \mathsf{E}[g(x_i,Y)|X=x_i]. \end{aligned} \tag{2.13}$$

which is the substitution law for conditional expectation. Note that if g in (2.13) is a function of Y only, then (2.13) reduces to (2.12). Also, if g is of product form, say $g(x,y)=h(x)k(y)$, then

$$\mathsf{E}[h(X)k(Y)|X=x_i] \;=\; h(x_i)\mathsf{E}[k(Y)|X=x_i].$$

Law of Total Probability for Expectation

In Section 2.4 we discussed the law of total probability, which shows how to compute probabilities in terms of conditional probabilities. We now derive the analogous formula for expectation. Write

$$\begin{aligned} \sum_i \mathsf{E}[g(X,Y)|X=x_i]\, p_X(x_i) &= \sum_i \left[\sum_j g(x_i,y_j)\, p_{Y|X}(y_j|x_i)\right] p_X(x_i) \\ &= \sum_i \sum_j g(x_i,y_j)\, p_{XY}(x_i,y_j) \\ &= \mathsf{E}[g(X,Y)]. \end{aligned}$$

In particular, if g is a function of Y only, then

$$\mathsf{E}[g(Y)] \;=\; \sum_i \mathsf{E}[g(Y)|X=x_i]\, p_X(x_i).$$

***Example* 2.36.** Light of intensity λ is directed at a photomultiplier that generates $X \sim \text{Poisson}(\lambda)$ primaries. The photomultiplier then generates Y secondaries, where given $X = n$, Y is conditionally geometric$_1\big((n+2)^{-1}\big)$. Find the expected number of secondaries and the correlation between the primaries and the secondaries.

Solution. The law of total probability for expectations says that

$$\mathsf{E}[Y] \;=\; \sum_{n=0}^{\infty} \mathsf{E}[Y|X=n]\, p_X(n),$$

where the range of summation follows because X is Poisson(λ). The next step is to compute the conditional expectation. The conditional pmf of Y is geometric$_1(p)$, where $p = (n+2)^{-1}$, and the mean of such a pmf is, by Problem 19, $1/(1 \Leftrightarrow p)$. Hence,

$$\mathsf{E}[Y] \;=\; \sum_{n=0}^{\infty} \left[1 + \frac{1}{n+1}\right] p_X(n) \;=\; \mathsf{E}\left[1 + \frac{1}{X+1}\right].$$

An easy calculation (Problem 14) shows that for $X \sim \text{Poisson}(\lambda)$,

$$\mathsf{E}\left[\frac{1}{X+1}\right] \;=\; [1 \Leftrightarrow e^{-\lambda}]/\lambda,$$

and so $\mathsf{E}[Y] = 1 + [1 \Leftrightarrow e^{-\lambda}]/\lambda$.

The correlation between X and Y is

$$\begin{aligned}
\mathsf{E}[XY] &= \sum_{n=0}^{\infty} \mathsf{E}[XY|X=n]\, p_X(n) \\
&= \sum_{n=0}^{\infty} n\, \mathsf{E}[Y|X=n]\, p_X(n) \\
&= \sum_{n=0}^{\infty} n \left[1 + \frac{1}{n+1}\right] p_X(n) \\
&= \mathsf{E}\left[X\left(1 + \frac{1}{X+1}\right)\right].
\end{aligned}$$

Now observe that

$$X\left(1 + \frac{1}{X+1}\right) \;=\; X + 1 \Leftrightarrow \frac{1}{X+1}.$$

It follows that

$$\mathsf{E}[XY] \;=\; \lambda + 1 \Leftrightarrow [1 \Leftrightarrow e^{-\lambda}]/\lambda.$$

2.6. Notes

Notes §2.1: Probabilities Involving Random Variables

***Note* 1.** According to Note 1 in Chapter 1, $\wp(A)$ is only defined for certain subsets $A \in \mathcal{A}$. Hence, in order that the probability

$$\wp(\{\omega \in \Omega : X(\omega) \in B\})$$

be defined, it is necessary that

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}. \tag{2.14}$$

To guarantee that this is true, it is convenient to consider $\wp(X \in B)$ only for sets B in some σ-field $\mathcal{B}$ of subsets of $\mathbb{R}$. The technical definition of a random variable is then as follows. A function X from Ω into $\mathbb{R}$ is a **random variable** if and only if (2.14) holds for every $B \in \mathcal{B}$. Usually $\mathcal{B}$ is usually taken to be the **Borel σ-field**; i.e., $\mathcal{B}$ is the smallest σ-field containing all the open subsets of $\mathbb{R}$. If $B \in \mathcal{B}$, then B is called a **Borel set**. It can be shown [4, pp. 182–183] that a real-valued function X satisfies (2.14) for all Borel sets B if and only if

$$\{\omega \in \Omega : X(\omega) \le x\} \in \mathcal{A}, \quad \text{for all } x \in \mathbb{R}.$$

***Note* 2.** In light of Note 1 above, we do not require that (2.1) hold for *all* sets B and C, but only for all *Borel* sets B and C.

***Note* 3.** We show that

$$\wp(X \in B, Y \in C) \;=\; \sum_i \sum_j I_B(x_i) I_C(y_j) p_{XY}(x_i, y_j).$$

Consider the disjoint events $\{Y = y_j\}$. Since $\sum_j \wp(Y = y_j) = 1$, we can use the law of total probability as in (1.13) with $A = \{X = x_i, Y \in C\}$ to write

$$\wp(X = x_i, Y \in C) \;=\; \sum_j \wp(X = x_i, Y \in C, Y = y_j).$$

Now observe that

$$\{Y \in C\} \cap \{Y = y_j\} \;=\; \begin{cases} \{Y = y_j\}, & y_j \in C, \\ \emptyset, & y_j \notin C. \end{cases}$$

Hence,

$$\wp(X = x_i, Y \in C) \;=\; \sum_j I_C(y_j) \wp(X = x_i, Y = y_j) \;=\; \sum_j I_C(y_j) p_{XY}(x_i, y_j).$$

The next step is to use (1.13) again, but this time with the disjoint events $\{X = x_i\}$ and $A = \{X \in B, Y \in C\}$. Then,

$$\wp(X \in B, Y \in C) \;=\; \sum_i \wp(X \in B, Y \in C, X = x_i).$$

Now observe that

$$\{X \in B\} \cap \{X = x_i\} \;=\; \begin{cases} \{X = x_i\}, & x_i \in B, \\ \emptyset, & x_i \notin B. \end{cases}$$

Hence,

$$\begin{aligned} \wp(X \in B, Y \in C) &= \sum_i I_B(x_i) \wp(X = x_i, Y \in C) \\ &= \sum_i I_B(x_i) \sum_j I_C(y_j) p_{XY}(x_i, y_j). \end{aligned}$$

Notes §2.2: Expectation

***Note* 4.** When z is complex,

$$\mathsf{E}[z^X] \;:=\; \mathsf{E}[\mathrm{Re}(z^X)] + j\mathsf{E}[\mathrm{Im}(z^X)].$$

By writing

$$z^n \;=\; r^n e^{jn\theta} \;=\; r^n[\cos(n\theta) + j\sin(n\theta)],$$

it is easy to check that

$$\mathsf{E}[z^X] \;=\; \sum_{n=0}^{\infty} z^n \wp(X = n).$$

Notes §2.4: Conditional Probability

***Note* 5.** Here is an alternative derivation of the fact that the sum of independent Bernoulli random variables is a binomial random variable. Let $X_1, X_2, \ldots$ be independent Bernoulli(p) random variables. Put

$$Y_n \;:=\; \sum_{i=1}^{n} X_i.$$

We need to show that $Y_n \sim$ binomial(n, p). The case $n = 1$ is trivial. Suppose the result is true for some $n \geq 1$. We show that it must be true for $n + 1$. Use the law of total probability to write

$$\wp(Y_{n+1} = k) \;=\; \sum_{i=0}^{n} \wp(Y_{n+1} = k | Y_n = i) \wp(Y_n = i). \tag{2.15}$$

To compute the conditional probability, we first observe that $Y_{n+1} = Y_n + X_{n+1}$. Also, since the X_i are independent, and since Y_n depends only on $X_1, \ldots, X_n$,

we see that Y_n and X_{n+1} are independent. Keeping this in mind, we apply the substitution law and write

$$\begin{aligned}\wp(Y_{n+1}=k|Y_n=i) &= \wp(Y_n+X_{n+1}=k|Y_n=i)\\ &= \wp(i+X_{n+1}=k|Y_n=i)\\ &= \wp(X_{n+1}=k \Leftrightarrow i|Y_n=i)\\ &= \wp(X_{n+1}=k \Leftrightarrow i).\end{aligned}$$

Since X_{n+1} takes only the values zero and one, this last probability is zero unless $i=k$ or $i=k \Leftrightarrow 1$. Returning to (2.15), we can write[§]

$$\wp(Y_{n+1}=k) \;=\; \sum_{i=k-1}^{k} \wp(X_{n+1}=k \Leftrightarrow i)\wp(Y_n=i).$$

Assuming that $Y_n \sim \text{binomial}(n,p)$, this becomes

$$\wp(Y_{n+1}=k) \;=\; p\binom{n}{k \Leftrightarrow 1}p^{k-1}(1 \Leftrightarrow p)^{n-(k-1)} + (1 \Leftrightarrow p)\binom{n}{k}p^k(1 \Leftrightarrow p)^{n-k}.$$

Using the easily verified identity,

$$\binom{n}{k \Leftrightarrow 1} + \binom{n}{k} \;=\; \binom{n+1}{k},$$

we see that $Y_{n+1} \sim \text{binomial}(n+1,p)$.

2.7. Problems

Problems §2.1: Probabilities Involving Random Variables

1. Let Y be an integer-valued random variable. Show that

$$\wp(Y=n) \;=\; \wp(Y>n \Leftrightarrow 1) \Leftrightarrow \wp(Y>n).$$

2. Let $X \sim \text{Poisson}(\lambda)$. Evaluate $\wp(X>1)$; your answer should be in terms of λ. Then compute the numerical value of $\wp(X>1)$ when $\lambda=1$. *Answer:* 0.264.

3. A certain photo-sensor fails to activate if it receives fewer than four photons in a certain time interval. If the number of photons is modeled by a Poisson(2) random variable X, find the probability that the sensor activates. *Answer:* 0.143.

4. Let $X \sim \text{geometric}_1(p)$.

[§]When $k=0$ or $k=n+1$, this sum actually has only one term, since $\wp(Y_n=-1) = \wp(Y_n=n+1)=0$.

(a) Show that $\wp(X > n) = p^n$.

(b) Compute $\wp(\{X > n+k\}|\{X > n\})$. *Hint:* If $A \subset B$, then $A \cap B = A$.

Remark. Your answer to (b) should not depend on n. For this reason, the geometric random variable is said to have the **memoryless property**. For example, let X model the number of the toss on which the first heads occurs in a sequence of coin tosses. Then given a heads has not occurred up to and including time n, the conditional probability that a heads does not occur in the next k tosses does not depend on n. In other words, given that no heads occurs on tosses $1, \ldots, n$ has no effect on the conditional probability of heads occurring in the future. Future tosses do not remember the past.

*5. From your solution of Problem 4(b), you can see that if $X \sim \text{geometric}_1(p)$, then $\wp(\{X > n + k\}|\{X > n\}) = \wp(X > k)$. Now prove the converse; i.e., show that if Y is a positive integer-valued random variable such that $\wp(\{Y > n + k\}|\{Y > n\}) = \wp(Y > k)$, then $Y \sim \text{geometric}_1(p)$, where $p = \wp(Y > 1)$. *Hint:* First show that $\wp(Y > n) = \wp(Y > 1)^n$; then apply Problem 1.

6. At the Chicago IRS office, there are m independent auditors. The kth auditor processes X_k tax returns per day, where X_k is Poisson distributed with parameter $\lambda > 0$. The office's performance is unsatisfactory if any auditor processes fewer than 2 tax returns per day. Find the probability that the office performance is unsatisfactory.

7. An astronomer has recently discovered n similar galaxies. For $i = 1, \ldots, n$, let X_i denote the number of black holes in the ith galaxy, and assume the X_i are independent Poisson(λ) random variables.

(a) Find the probability that at least one of the galaxies contains two or more black holes.

(b) Find the probability that all n galaxies have at least one black hole.

(c) Find the probability that all n galaxies have exactly one black hole.

Your answers should be in terms of n and λ.

8. There are 29 stocks on the Get Rich Quick Stock Exchange. The price of each stock (in whole dollars) is $\text{geometric}_0(p)$ (same p for all stocks). Prices of different stocks are independent. If $p = 0.7$, find the probability that at least one stock costs more than 10 dollars. *Answer:* 0.44.

9. Let $X_1, \ldots, X_n$ be independent, $\text{geometric}_1(p)$ random variables. Evaluate $\wp(\min(X_1, \ldots, X_n) > \ell)$ and $\wp(\max(X_1, \ldots, X_n) \leq \ell)$.

10. A class consists of 15 students. Each student has probability $p = 0.1$ of getting an "A" in the course. Find the probability that exactly one

student receives an "A." Assume the students' grades are independent. *Answer:* 0.343.

11. In a certain lottery game, the player chooses three digits. The player wins if at least two out of three digits match the random drawing for that day in both position and value. Find the probability that the player wins. Assume that the digits of the random drawing are independent and equally likely. *Answer:* 0.028.

12. Blocks on a computer disk are good with probability p and faulty with probability $1 \Leftrightarrow p$. Blocks are good or bad independently of each other. Let Y denote the location (starting from 1) of the first bad block. Find the pmf of Y.

13. Let X and Y be jointly discrete, integer-valued random variables with joint pmf

$$p_{XY}(i,j) \;=\; \begin{cases} \dfrac{3^{j-1}e^{-3}}{j!}, & i = 1,\ j \geq 0, \\ 4\dfrac{6^{j-1}e^{-6}}{j!}, & i = 2,\ j \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find the marginal pmfs $p_X(i)$ and $p_Y(j)$, and determine whether or not X and Y are independent.

Problems §2.2: Expectation

14. If X is Poisson(λ), compute $\mathsf{E}[1/(X+1)]$.

15. A random variable X has mean 2 and variance 7. Find $\mathsf{E}[X^2]$.

16. Let X be a random variable with mean m and variance σ^2. Find the constant c that best approximates the random variable X in the sense that c minimizes the **mean squared error** $\mathsf{E}[(X \Leftrightarrow c)^2]$.

17. Find $\mathsf{var}(X)$ if X has probability generating function

$$G_X(z) \;=\; \tfrac{1}{6} + \tfrac{1}{6}z + \tfrac{2}{3}z^2.$$

18. Find $\mathsf{var}(X)$ if X has probability generating function

$$G_X(z) \;=\; \left(\frac{2+z}{3}\right)^5.$$

19. Evaluate $G_X(z)$ for the cases $X \sim \text{geometric}_0(p)$ and $X \sim \text{geometric}_1(p)$. Use your results to find the mean and variance of X in each case.

20. The probability generating function of $Y \sim \text{binomial}(n,p)$ was found in Example 2.20. Use it to find the mean and variance of Y.

21. Compute $\mathsf{E}[(X+Y)^3]$ if X and Y are independent Bernoulli(p) random variables.

22. Let $X_1, \ldots, X_n$ be independent Poisson(λ) random variables. Find the probability generating function of $Y := X_1 + \cdots + X_n$.

23. For $i = 1, \ldots, n$, let $X_i \sim \text{Poisson}(\lambda_i)$. Put

$$Y \;:=\; \sum_{i=1}^{n} X_i.$$

Find $\wp(Y = 2)$ if the X_i are independent.

24. A certain digital communication link has bit-error probability p. In a transmission of n bits, find the probability that k bits are received incorrectly, assuming bit errors occur independently.

25. A new school has M classrooms. For $i = 1, \ldots, M$, let n_i denote the number of seats in the ith classroom. Suppose that the number of students in the ith classroom is binomial(n_i, p) and independent. Let Y denote the total number of students in the school. Find $\wp(Y = k)$.

26. Let $X_1, \ldots, X_n$ be i.i.d. with $\wp(X_i = 1) = 1 \Leftrightarrow p$ and $\wp(X_i = 2) = p$. If $Y := X_1 + \cdots + X_n$, find $\wp(Y = k)$ for all k.

27. Ten-bit codewords are transmitted over a noisy channel. Bits are flipped independently with probability p. If no more than two bits of a codeword are flipped, the codeword can be correctly decoded. Find the probability that a codeword cannot be correctly decoded.

28. From a well-shuffled deck of 52 playing cards you are dealt 14 cards. What is the probability that 2 cards are spades, 3 are hearts, 4 are diamonds, and 5 are clubs? *Answer:* 0.0116.

29. From a well-shuffled deck of 52 playing cards you are dealt 5 cards. What is the probability that all 5 cards are of the same suit? *Answer:* 0.00198.

*30. A generalization of the binomial coefficient is the **multinomial coefficient**. Let $m_1, \ldots, m_r$ be nonnegative integers that sum to n. Then

$$\binom{n}{m_1, \ldots, m_r} \;:=\; \frac{n!}{m_1! \cdots m_r!}.$$

If words of length n are composed of r-ary symbols, then the multinomial coefficient is the number of words with exactly m_1 symbols of type 1, m_2 symbols of type 2, ..., and m_r symbols of type r. The **multinomial theorem** says that

$$\left(\sum_{j=1}^{J} a_j \right)^{k_J}$$

is equal to

$$\sum_{k_{J-1}=0}^{k_J} \cdots \sum_{k_1=0}^{k_2} \binom{k_J}{k_1, k_2 - k_1, \ldots, k_J - k_{J-1}} a_1^{k_1} a_2^{k_2-k_1} \cdots a_J^{k_J - k_{J-1}}.$$

Derive the multinomial theorem by induction on J.

Remark. The multinomial theorem can also be expressed in the form

$$\left(\sum_{j=1}^{J} a_j\right)^n = \sum_{m_1+\cdots+m_J=n} \binom{n}{m_1, m_2, \ldots, m_J} a_1^{m_1} a_2^{m_2} \cdots a_J^{m_J}.$$

31. Make a table comparing both sides of the Poisson approximation of binomial probabilities,

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{(np)^k e^{-np}}{k!}, \quad n \text{ large, } p \text{ small,}$$

for $k = 0, 1, 2, 3, 4, 5$ if $n = 100$ and $p = 1/100$. *Hint:* If MATLAB is available, the binomial probability can be written

$$\texttt{nchoosek(n,k)} * \texttt{p\^{}k} * \texttt{(1-p)\^{}(n-k)}$$

and the Poisson probability can be written

$$\texttt{(n*p)\^{}k} * \texttt{exp(-n*p)/factorial(k)}.$$

32. *Betting on Fair Games.* Let $X \sim$ Bernoulli(p) random variable. For example, we could let $X = 1$ model the result of a coin toss being heads. Or we could let $X = 1$ model your winning the lottery. In general, a bettor wagers a stake of s dollars that $X = 1$ with a bookmaker who agrees to pay d dollars to the bettor if $X = 1$ occurs; if $X = 0$, the stake s is kept by the bookmaker. Thus, the net income of the bettor is

$$Y := dX - s(1-X),$$

since if $X = 1$, the bettor receives $Y = d$ dollars, and if $X = 0$, the bettor receives $Y = -s$ dollars; i.e., loses s dollars. Of course the net income to the bookmaker is $-Y$. If the wager is fair to both the bettor and the bookmaker, then we should have $\mathsf{E}[Y] = 0$. In other words, on average, the net income to either party is zero. Show that a fair wager requires that $d/s = (1-p)/p$.

33. *Odds.* Let $X \sim$ Bernoulli(p) random variable. We say that the (fair) **odds** *against* $X = 1$ are n_2 to n_1 (written $n_2 : n_1$) if n_2 and n_1 are positive integers satisfying $n_2/n_1 = (1-p)/p$. Typically, n_2 and n_1 are chosen to

have no common factors. Conversely, we say that the odds *for* $X = 1$ are n_1 to n_2 if $n_1/n_2 = p/(1 \Leftrightarrow p)$. Consider a state lottery game in which players wager one dollar that they can correctly guess a randomly selected three-digit number in the range 000–999. The state offers a payoff of \$500 for a correct guess.

(a) What is the probability of correctly guessing the number?

(b) What are the (fair) odds against guessing correctly?

(c) The odds against actually offered by the state are determined by the ratio of the payoff divided by the stake, in this case, 500 : 1. Is the game fair to the bettor? If not, what should the payoff be to make it fair? (See the preceding problem for the notion of "fair.")

*34. These results are needed for Examples 2.14 and 2.15. Show that

$$\sum_{k=1}^{\infty} \frac{1}{k} \;=\; \infty$$

and that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \;<\; 2.$$

(The exact value of the second sum is $\pi^2/6$ [38, p. 198].) *Hints:* For the first formula, use the inequality

$$\int_k^{k+1} \frac{1}{t}\,dt \;\leq\; \int_k^{k+1} \frac{1}{k}\,dt \;=\; \frac{1}{k}.$$

For the second formula, use the inequality

$$\int_k^{k+1} \frac{1}{t^2}\,dt \;\geq\; \int_k^{k+1} \frac{1}{(k+1)^2}\,dt \;=\; \frac{1}{(k+1)^2}.$$

35. Let X be a discrete random variable taking finitely many distinct values $x_1, \ldots, x_n$. Let $p_i := \wp(X = x_i)$ be the corresponding probability mass function. Consider the function

$$g(x) \;:=\; \Leftrightarrow \log \wp(X = x).$$

Observe that $g(x_i) = \Leftrightarrow \log p_i$. The **entropy** of X is defined by

$$H(X) := \mathsf{E}[g(X)] = \sum_{i=1}^{n} g(x_i)\wp(X = x_i) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}.$$

If all outcomes are equally likely, i.e., $p_i = 1/n$, find $H(X)$. If X is a constant random variable, i.e., $p_j = 1$ for some j and $p_i = 0$ for $i \neq j$, find $H(X)$.

*36. *Jensen's inequality.* Recall that a real-valued function g defined on an interval I is **convex** if for all $x, y \in I$ and all $0 \le \lambda \le 1$,

$$g\big(\lambda x + (1 \Leftrightarrow \lambda)y\big) \;\le\; \lambda g(x) + (1 \Leftrightarrow \lambda)g(y).$$

Let g be a convex function, and let X be a discrete random variable taking finitely many values, say n values, all in I. Derive **Jensen's inequality**,

$$\mathsf{E}[g(X)] \;\ge\; g(\mathsf{E}[X]).$$

Hint: Use induction on n.

*37. Derive **Lyapunov's inequality**,

$$\mathsf{E}[|Z|^\alpha]^{1/\alpha} \;\le\; \mathsf{E}[|Z|^\beta]^{1/\beta}, \quad 1 \le \alpha < \beta < \infty.$$

Hint: Apply Jensen's inequality to the convex function $g(x) = x^{\beta/\alpha}$ and the random variable $X = |Z|^\alpha$.

*38. A discrete random variable is said to be nonnegative, denoted by $X \ge 0$, if $\wp(X \ge 0) = 1$; i.e., if

$$\sum_i I_{[0,\infty)}(x_i)\wp(X = x_i) \;=\; 1.$$

(a) Show that for a nonnegative random variable, if $x_k < 0$ for some k, then $\wp(X = x_k) = 0$.

(b) Show that for a nonnegative random variable, $\mathsf{E}[X] \ge 0$.

(c) If X and Y are discrete random variables, we write $X \ge Y$ if $X \Leftrightarrow Y \ge 0$. Show that if $X \ge Y$, then $\mathsf{E}[X] \ge \mathsf{E}[Y]$; i.e., expectation is **monotone**.

Problems §2.3: The Weak Law of Large Numbers

39. Show that $\mathsf{E}[M_n] = m$. Also show that for any constant c, $\mathsf{var}(cX) = c^2\,\mathsf{var}(X)$.

40. Let $X_1, X_2, \ldots, X_n$ be i.i.d. geometric$_1(p)$ random variables, and put $Y := X_1 + \cdots + X_n$. Find $\mathsf{E}[Y]$, $\mathsf{var}(Y)$, and $\mathsf{E}[Y^2]$. Also find the moment generating function of Y.

Remark. We say that Y is a **negative binomial** or **Pascal** random variable with parameters n and p.

41. Show by counterexample that being uncorrelated does not imply independence. *Hint:* Let $\wp(X = \pm 1) = \wp(X = \pm 2) = 1/4$, and put $Y := |X|$. Show that $\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y]$, but $\wp(X = 1, Y = 1) \ne \wp(X = 1)\,\wp(Y = 1)$.

42. Let $X \sim \text{geometric}_0(1/4)$. Compute both sides of Markov's inequality,

$$\wp(X \geq 2) \;\leq\; \frac{\mathsf{E}[X]}{2}.$$

43. Let $X \sim \text{geometric}_0(1/4)$. Compute both sides of Chebyshev's inequality,

$$\wp(X \geq 2) \;\leq\; \frac{\mathsf{E}[X^2]}{4}.$$

44. Student heights range from 120 cm to 220 cm. To estimate the average height, determine how many students' heights should be measured to make the sample mean within 0.25 cm of the true mean height with probability at least 0.9. Assume measurements are uncorrelated and have variance $\sigma^2 = 1$. What if you only want to be within 1 cm of the true mean height with probability at least 0.9?

45. Show that for an arbitrary sequence of random variables, it is not always true that for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \wp(|M_n \Leftrightarrow m| \geq \varepsilon) \;=\; 0.$$

Hint: Let Z be a nonconstant random variable and put $X_i := Z$ for $i = 1, 2, \ldots$. To be specific, try $Z \sim \text{Bernoulli}(1/2)$ and $\varepsilon = 1/4$.

46. Let X and Y be two random variables with means m_X and m_Y and variances σ_X^2 and σ_Y^2. The **correlation coefficient** of X and Y is

$$\rho \;:=\; \frac{\mathsf{E}[(X \Leftrightarrow m_X)(Y \Leftrightarrow m_Y)]}{\sigma_X \sigma_Y}.$$

Note that $\rho = 0$ if and only if X and Y are uncorrelated. Suppose X cannot be observed, but we are able to measure Y. We wish to estimate X by using the quantity aY, where a is a suitable constant. Assuming $m_X = m_Y = 0$, find the constant a that minimizes the mean squared error $\mathsf{E}[(X \Leftrightarrow aY)^2]$. Your answer should depend on σ_X, σ_Y and ρ.

Problems §2.4: Conditional Probability

47. Let X and Y be integer-valued random variables. Suppose that conditioned on $X = i$, $Y \sim \text{binomial}(n, p_i)$, where $0 < p_i < 1$. Evaluate $\wp(Y < 2|X = i)$.

48. Let X and Y be integer-valued random variables. Suppose that conditioned on $Y = j$, $X \sim \text{Poisson}(\lambda_j)$. Evaluate $\wp(X > 2|Y = j)$.

49. Let X and Y be independent random variables. Show that $p_{X|Y}(x_i|y_j) = p_X(x_i)$ and $p_{Y|X}(y_j|x_i) = p_Y(y_j)$.

50. Let X and Y be independent with $X \sim \text{geometric}_0(p)$ and $Y \sim \text{geometric}_0(q)$. Put $T := X \Leftrightarrow Y$, and find $\wp(T = n)$ for all n.

51. When a binary optical communication system transmits a 1, the receiver output is a Poisson(μ) random variable. When a 2 is transmitted, the receiver output is a Poisson(ν) random variable. Given that the receiver output is equal to 2, find the conditional probability that a 1 was sent. Assume messages are equally likely.

52. In a binary communication system, when a 0 is sent, the receiver outputs a random variable Y that is $\text{geometric}_0(p)$. When a 1 is sent, the receiver output $Y \sim \text{geometric}_0(q)$, where $q \neq p$. Given that the receiver output $Y = k$, find the conditional probability that the message sent was a 1. Assume messages are equally likely.

53. Apple crates are supposed to contain only red apples, but occasionally a few green apples are found. Assume that the number of red apples and the number of green apples are independent Poisson random variables with parameters ρ and γ, respectively. Given that a crate contains a total of k apples, find the conditional probability that none of the apples is green.

54. Let $X \sim \text{Poisson}(\lambda)$, and suppose that given $X = n$, $Y \sim \text{Bernoulli}(1/(n+1))$. Find $\wp(X = n|Y = 1)$.

55. Let $X \sim \text{Poisson}(\lambda)$, and suppose that given $X = n$, $Y \sim \text{binomial}(n, p)$. Find $\wp(X = n|Y = k)$ for $n \geq k$.

56. Let X and Y be independent binomial(n, p) random variables. Find the conditional probability that $X > k$ given that $\max(X, Y) > k$ if $n = 100$, $p = 0.01$, and $k = 1$. *Answer:* 0.284.

57. Let $X \sim \text{geometric}_0(p)$ and $Y \sim \text{geometric}_0(q)$, and assume X and Y are independent.

    (a) Find $\wp(XY = 4)$.

    (b) Put $Z := X+Y$ and find $p_Z(j)$ for all j using the discrete convolution formula (2.11). Treat the cases $p \neq q$ and $p = q$ separately.

58. Let X and Y be independent random variables, each taking the values $0, 1, 2, 3$ with equal probability. Put $Z := X + Y$ and sketch $p_Z(j)$ for all j. *Hint:* Use the discrete convolution formula (2.11), and use graphical convolution techniques.

Problems §2.5: Conditional Expectation

59. Let X and Y be as in Problem 13. Compute $\mathsf{E}[Y|X = i]$, $\mathsf{E}[Y]$, and $\mathsf{E}[X|Y = j]$.

60. Let X and Y be as in Example 2.30. Find $\mathsf{E}[Y]$, $\mathsf{E}[XY]$, $\mathsf{E}[Y^2]$, and $\mathsf{var}(Y)$.

61. Let X and Y be as in Example 2.31. Find $\mathsf{E}[Y|X=1]$, $\mathsf{E}[Y|X=0]$, $\mathsf{E}[Y]$, $\mathsf{E}[Y^2]$, and $\mathsf{var}(Y)$.

62. Let $X \sim \text{Bernoulli}(2/3)$, and suppose that given $X = i$, $Y \sim \text{Poisson}\big(3(i+1)\big)$. Find $\mathsf{E}[(X+1)Y^2]$.

63. Let $X \sim \text{Poisson}(\lambda)$, and suppose that given $X = n$, $Y \sim \text{Bernoulli}(1/(n+1))$. Find $\mathsf{E}[XY]$.

64. Let $X \sim \text{geometric}_1(p)$, and suppose that given $X = n$, $Y \sim \text{Pascal}(n,q)$. Find $\mathsf{E}[XY]$.

65. Let X and Y be integer-valued random variables. Suppose that given $Y = k$, X is conditionally Poisson with parameter k. If Y has mean m and variance r, find $\mathsf{E}[X^2]$.

66. Let X and Y be independent random variables, with $X \sim \text{binomial}(n,p)$, and let $Y \sim \text{binomial}(m,p)$. Put $V := X + Y$. Find the pmf of V. Find $\wp(V = 10|X = 4)$ (assume $n \geq 4$ and $m \geq 6$).

67. Let X and Y be as in Problem 13. Find the probability generating function of Y, and then find $\wp(Y = 2)$.

68. Let X and Y be as in Example 2.30. Find $G_Y(z)$.

CHAPTER 3

Continuous Random Variables

In Chapter 2, the only specific random variables we considered were discrete such as the Bernoulli, binomial, Poisson, and geometric. In this chapter we consider a class of random variables that are allowed to take on a continuum of values. These random variables are called continuous random variables and are introduced in Section 3.1. Their expectation is defined in Section 3.2 and used to develop the concepts of moment generating function (Laplace transform) and characteristic function (Fourier transform). In Section 3.3 expectation of multiple random variables is considered. Applications of characteristic functions to sums of independent random variables are illustrated. In Section 3.4 Markov's inequality, Chebyshev's inequality, and the Chernoff bound illustrate simple techniques for bounding probabilities in terms of expectations.

3.1. Definition and Notation

Recall that a real-valued function $X(\omega)$ defined on a sample space Ω is called a random variable. We say that X is a **continuous random variable** if $\wp(X \in B)$ has the form

$$\wp(X \in B) \;=\; \int_B f(t)\,dt \;:=\; \int_{-\infty}^{\infty} I_B(t) f(t)\,dt \tag{3.1}$$

for some integrable function f. Since $\wp(X \in \mathbb{R}) = 1$, f must integrate to one; i.e., $\int_{-\infty}^{\infty} f(t)\,dt = 1$. Further, since $\wp(X \in B) \ge 0$ for all B, it can be shown that f must be nonnegative.[1] A nonnegative function that integrates to one is called a **probability density function** (pdf).

Here are some examples of continuous random variables. A summary of the more common ones can be found on the inside of the back cover.

The simplest continuous random variable is the **uniform**. It is used to model experiments in which the outcome is constrained to lie in a known interval, say $[a, b]$, and all outcomes are equally likely. We used this model in the bus Example 1.10 in Chapter 1. We write $f \sim \text{uniform}[a, b]$ if $a < b$ and

$$f(x) \;=\; \begin{cases} \dfrac{1}{b \Leftrightarrow a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

To verify that f integrates to one, first note that since $f(x) = 0$ for $x < a$ and $x > b$, we can write

$$\int_{-\infty}^{\infty} f(x)\,dx \;=\; \int_a^b f(x)\,dx.$$

Next, for $a \le x \le b$, $f(x) = 1/(b \Leftrightarrow a)$, and so

$$\int_a^b f(x)\,dx \;=\; \int_a^b \frac{1}{b \Leftrightarrow a}\,dx \;=\; 1.$$

This calculation illustrates a very important technique that is often incorrectly carried out by beginning students: *First* modify the limits of integration, *then* substitute the appropriate formula for $f(x)$. For example, it is quite common to see the *incorrect* calculation,

$$\int_{-\infty}^{\infty} f(x)\,dx \;=\; \int_{-\infty}^{\infty} \frac{1}{b \Leftrightarrow a}\,dx \;=\; \infty.$$

***Example* 3.1.** If X is a continuous random variable with density $f \sim$ uniform$[\Leftrightarrow 1, 3]$, find $\wp(X \le 0)$, $\wp(X \le 1)$, and $\wp(X > 1 | X > 0)$.

Solution. To begin, write

$$\wp(X \le 0) \;=\; \wp(X \in (\Leftrightarrow\infty, 0]) \;=\; \int_{-\infty}^{0} f(x)\,dx.$$

Since $f(x) = 0$ for $x < \Leftrightarrow 1$, we can modify the limits of integration and obtain

$$\wp(X \le 0) \;=\; \int_{-1}^{0} f(x)\,dx \;=\; \int_{-1}^{0} \frac{1}{4}\,dx \;=\; 1/4.$$

Similarly, $\wp(X \le 1) = \int_{-1}^{1} 1/4\,dx = 1/2$. To calculate

$$\wp(X > 1 | X > 0) \;=\; \frac{\wp(\{X > 1\} \cap \{X > 0\})}{\wp(X > 0)},$$

observe that the denominator is simply $\wp(X > 0) = 1 \Leftrightarrow \wp(X \le 0) = 1 \Leftrightarrow 1/4 = 3/4$. As for the numerator,

$$\wp(\{X > 1\} \cap \{X > 0\}) = \wp(X > 1) = 1 \Leftrightarrow \wp(X \le 1) = 1/2.$$

Thus, $\wp(X > 1 | X > 0) = (1/2)/(3/4) = 2/3$.

Another simple continuous random variable is the **exponential** with parameter $\lambda > 0$. We write $f \sim \exp(\lambda)$ if

$$f(x) \;=\; \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

It is easy to check that f integrates to one. The exponential random variable is often used to model lifetimes, such as how long a lightbulb burns or how long it takes for a computer to transmit a message from one node to another.

The exponential random variable also arises as a function of other random variables. For example, in Problem 35 you will show that if $U \sim \text{uniform}(0,1)$, then $X = \ln(1/U)$ is $\exp(1)$. We also point out that if U and V are independent Gaussian random variables,* then $U^2 + V^2$ is exponential^2 and $\sqrt{U^2+V^2}$ is Rayleigh (defined in Problem 30).

Related to the exponential is the **Laplace**, sometimes called the **double-sided exponential**. For $\lambda > 0$, we write $f \sim \text{Laplace}(\lambda)$ if

$$f(x) \;=\; \tfrac{\lambda}{2} e^{-\lambda|x|}.$$

You will show in Problem 45 that the difference of two independent exponential random variables is a Laplace random variable.

***Example* 3.2.** If X is a continuous random variable with density $f \sim \text{Laplace}(\lambda)$, find

$$\wp(\text{⇔}3 \le X \le \text{⇔}2 \text{ or } 0 \le X \le 3).$$

Solution. The desired probability can be written as

$$\wp(\{\text{⇔}3 \le X \le \text{⇔}2\} \cup \{0 \le X \le 3\}).$$

Since these are disjoint events, the probability of the union is the sum of the individual probabilities. We therefore need to compute

$$\wp(\text{⇔}3 \le X \le \text{⇔}2) \;=\; \int_{-3}^{-2} \tfrac{\lambda}{2} e^{-\lambda|x|}\,dx \;=\; \tfrac{\lambda}{2}\int_{-3}^{-2} e^{\lambda x}\,dx,$$

which is equal to $(e^{-2\lambda} \text{⇔} e^{-3\lambda})/2$, and

$$\wp(0 \le X \le 3) \;=\; \int_{0}^{3} \tfrac{\lambda}{2} e^{-\lambda|x|}\,dx \;=\; \tfrac{\lambda}{2}\int_{0}^{3} e^{-\lambda x}\,dx,$$

which is equal to $(1 \text{⇔} e^{-3\lambda})/2$. The desired probability is then

$$\frac{1 \text{⇔} 2e^{-3\lambda} + e^{-2\lambda}}{2}.$$

The **Cauchy** random variable with parameter $\lambda > 0$ is also easy to work with. We write $f \sim \text{Cauchy}(\lambda)$ if

$$f(x) \;=\; \frac{\lambda/\pi}{\lambda^2 + x^2}.$$

Since $(1/\pi)(d/dx)\tan^{-1}(x/\lambda) = f(x)$, and since $\tan^{-1}(\infty) = \pi/2$, it is easy to check that f integrates to one. The Cauchy random variable arises as the quotient of independent Gaussian random variables (Problem 18 in Chapter 5).

*Gaussian random variables are defined later in this section.

***Example* 3.3.** In the λ-lottery you choose a number λ with $1 \le \lambda \le 10$. Then a random variable X is chosen according to the Cauchy density with parameter λ. If $|X| \ge 1$, then you win the lottery. Which value of λ should you choose to maximize your probability of winning?

Solution. Your probability of winning is

$$\begin{aligned} \wp(|X| \ge 1) &= \wp(X \ge 1 \text{ or } X \le \Leftrightarrow 1) \\ &= \int_1^\infty f(x)\,dx + \int_{-\infty}^1 f(x)\,dx, \end{aligned}$$

where $f(x) = (\lambda/\pi)/(\lambda^2 + x^2)$ is the Cauchy density. Since the Cauchy density is an even function,

$$\wp(|X| \ge 1) = 2\int_1^\infty \frac{\lambda/\pi}{\lambda^2 + x^2}\,dx.$$

Now make the change of variable $y = x/\lambda$, $dy = dx/\lambda$, to get

$$\wp(|X| \ge 1) = 2\int_{1/\lambda}^\infty \frac{1/\pi}{1 + y^2}\,dy.$$

Since the integrand is nonnegative, the integral is maximized by minimizing $1/\lambda$ or by maximizing λ. Hence, choosing $\lambda = 10$ maximizes your probability of winning.

The most important density is the **Gaussian** or **normal**. As a consequence of the central limit theorem, whose discussion is taken up in Chapter 4, Gaussian random variables are good models for sums of many independent random variables. Hence, Gaussian random variables often arise as noise models in communication and control systems. For $\sigma^2 > 0$, we write $f \sim N(m, \sigma^2)$ if

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\Leftrightarrow \frac{1}{2}\left(\frac{x \Leftrightarrow m}{\sigma}\right)^2\right],$$

where σ is the positive square root of σ^2. If $m = 0$ and $\sigma^2 = 1$, we say that f is a **standard normal density**. A graph of the standard normal density is shown in Figure 3.1.

To verify that an arbitrary normal density integrates to one, we proceed as follows. (For an alternative derivation, see Problem 14.) First, making the change of variable $t = (x \Leftrightarrow m)/\sigma$ shows that

$$\int_{-\infty}^\infty f(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-t^2/2}\,dt.$$

So, without loss of generality, we may assume f is a standard normal density with $m = 0$ and $\sigma = 1$. We then need to show that $I := \int_{-\infty}^\infty e^{-x^2/2}\,dx = \sqrt{2\pi}$.

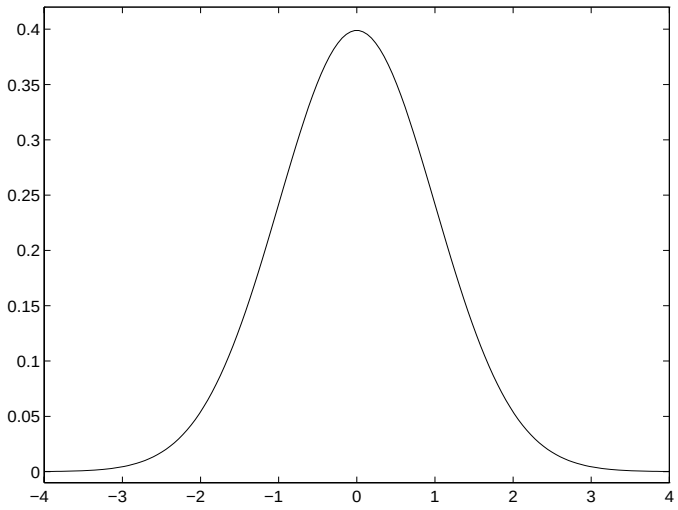

Figure 3.1. Standard normal density.

The trick is to show instead that $I^2 = 2\pi$. First write

$$I^2 \;=\; \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,dy\right).$$

Now write the product of integrals as the iterated integral

$$I^2 \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy.$$

Next, we interpret this as a double integral over the whole plane and change to polar coordinates. This yields

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,dr\,d\theta \\ &= \int_0^{2\pi}\left(\Leftrightarrow e^{-r^2/2}\,\Big|_0^{\infty}\right) d\theta \\ &= 2\pi. \end{aligned}$$

***Example* 3.4.** If f denotes the standard normal density, show that

$$\int_{-\infty}^{0} f(x)\,dx \;=\; \int_0^{\infty} f(x)\,dx \;=\; 1/2.$$

Solution. Since $f(x) = e^{-x^2/2}/\sqrt{2\pi}$ is an even function of x, the two integrals are equal. Since the sum of the two integrals is $\int_{-\infty}^{\infty} f(x)\,dx = 1$, each individual integral must be $1/2$.

The Paradox of Continuous Random Variables

Let $X \sim \text{uniform}(0,1)$, and fix any point $x_0 \in (0,1)$. What is $\wp(X = x_0)$? Consider the interval $[x_0 \Leftrightarrow 1/n, x_0 + 1/n]$. Write

$$\wp(X \in [x_0 \Leftrightarrow 1/n, x_0 + 1/n]) \;=\; \int_{x_0 - 1/n}^{x_0 + 1/n} 1\,dx \;=\; 2/n \;\to\; 0$$

as $n \to \infty$. Since the interval $[x_0 \Leftrightarrow 1/n, x_0 + 1/n]$ converges to the singleton set $\{x_0\}$, we conclude[†] that $\wp(X = x_0) = 0$ for every $x_0 \in (0,1)$. We are thus confronted with the fact that continuous random variables take no fixed value with positive probability! The way to understand this apparent paradox is to realize that continuous random variables are an *idealized model* of what we normally think of as continuous-valued measurements. For example, a voltmeter only shows a certain number of digits after the decimal point, say 5.127 volts because physical devices have limited precision. Hence, the measurement $X = 5.127$ should be understood as saying that

$$5.1265 \;\le\; X \;<\; 5.1275,$$

since all numbers in this range round to 5.127. Now there is no paradox because $\wp(5.1265 \le X < 5.1275)$ has positive probability.

You may still ask, "Why not just use a discrete random variable taking the distinct values $k/1000$, where k is any integer?" After all, this would model the voltmeter in question. The answer is that if you get a better voltmeter, you need to redefine the random variable, while with the idealized, continuous-random-variable model, even if the voltmeter changes, the random variable does not.

Remark. If B is any set with finitely many points, or even countably many points, then $\wp(X \in B) = 0$ when X is a continuous random variable. To see this, suppose $B = \{x_1, x_2, \ldots\}$ where the x_i are distinct real numbers. Then

$$\wp(X \in B) \;=\; \wp\Big(\bigcup_{i=1}^{\infty} \{x_i\}\Big) \;=\; \sum_{i=1}^{\infty} \wp(X = x_i) \;=\; 0,$$

since, as argued above, each term is zero.

3.2. Expectation of a Single Random Variable

Let X be a continuous random variable with density f, and let g be a real-valued function taking finitely many distinct values $y_j \in \mathbb{R}$. We claim that

$$\mathsf{E}[g(X)] \;=\; \int_{-\infty}^{\infty} g(x) f(x)\,dx. \tag{3.2}$$

[†] This argument is made rigorously for arbitrary continuous random variables in Section 4.6.

First note that the assumptions on g imply $Y := g(X)$ is a discrete random variable whose expectation is defined as

$$\mathsf{E}[Y] \;=\; \sum_j y_j \, \wp(Y = y_j).$$

Let $B_j := \{x : g(x) = y_j\}$. Observe that $X \in B_j$ if and only if $g(X) = y_j$, or equivalently, if and only if $Y = y_j$. Hence,

$$\mathsf{E}[Y] \;=\; \sum_j y_j \, \wp(X \in B_j).$$

Since X is a continuous random variable,

$$\wp(X \in B_j) \;=\; \int_{-\infty}^{\infty} I_{B_j}(x) f(x)\, dx.$$

Substituting this into the preceding equation, and writing $g(X)$ instead of Y, we have

$$\mathsf{E}[g(X)] \;=\; \sum_j y_j \int_{-\infty}^{\infty} I_{B_j}(x) f(x)\, dx \;=\; \int_{-\infty}^{\infty} \left[\sum_j y_j I_{B_j}(x)\right] f(x)\, dx.$$

It suffices to show that the sum in brackets is exactly $g(x)$. Fix any $x \in \mathbb{R}$. Then $g(x) = y_k$ for some k. In other words, $x \in B_k$. Now the B_j are disjoint because the y_j are distinct. Hence, for $x \in B_k$, the sum in brackets reduces to y_k, which is exactly $g(x)$.

We would like to apply (3.2) for more arbitrary functions g. However, if $g(X)$ is not a discrete random variable, its expectation has not yet been defined! This raises the question of how to define the expectation of an arbitrary random variable. The approach is to approximate X by a sequence of discrete random variables (for which expectation was defined in Chapter 2) and then define $\mathsf{E}[X]$ to be the limit of the expectations of the approximations. To be more precise about this, consider the sequence of functions q_n sketched in Figure 3.2. Since each q_n takes only finitely many distinct values, $q_n(X)$ is a discrete random variable for which $\mathsf{E}[q_n(X)]$ is defined. Since $q_n(X) \to X$, we then *define*[3] $\mathsf{E}[X] := \lim_{n\to\infty} \mathsf{E}[q_n(X)]$.

Now suppose X is a continuous random variable. Since $q_n(X)$ is a discrete random variable, (3.2) applies, and we can write

$$\mathsf{E}[X] \;:=\; \lim_{n\to\infty} \mathsf{E}[q_n(X)] \;=\; \lim_{n\to\infty} \int_{-\infty}^{\infty} q_n(x) f(x)\, dx.$$

Now bring the limit inside the integral,[4] and then use the fact that $q_n(x) \to x$. This yields

$$\mathsf{E}[X] \;=\; \int_{-\infty}^{\infty} \lim_{n\to\infty} q_n(x) f(x)\, dx \;=\; \int_{-\infty}^{\infty} x f(x)\, dx.$$

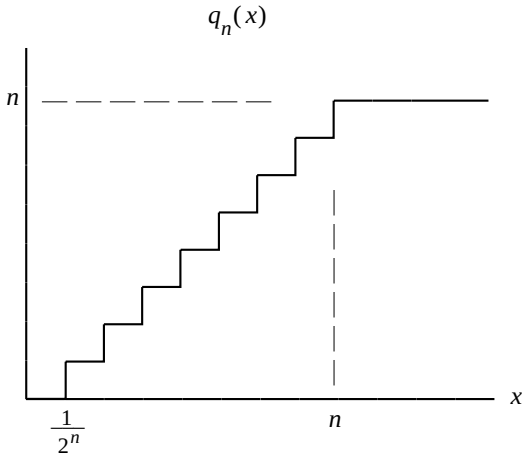

Figure 3.2. Finite-step quantizer $q_n(x)$ for approximating arbitrary random variables by discrete random variables. The number of steps is $n2^n$. In the figure, $n = 2$ and so there are 8 steps.

The same technique can be used to show that (3.2) holds even if g takes more than finitely many values. Write

$$\begin{aligned}
\mathsf{E}[g(X)] &:= \lim_{n\to\infty} \mathsf{E}[q_n(g(X))] \\
&= \lim_{n\to\infty} \int_{-\infty}^{\infty} q_n(g(x)) f(x)\, dx \\
&= \int_{-\infty}^{\infty} \lim_{n\to\infty} q_n(g(x)) f(x)\, dx \\
&= \int_{-\infty}^{\infty} g(x) f(x)\, dx.
\end{aligned}$$

***Example* 3.5.** If $X \sim$ uniform$[a,b]$ random variable, find $\mathsf{E}[X]$ and $\mathsf{E}[\cos(X)]$.

Solution. To find $\mathsf{E}[X]$, write

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{1}{b \ominus a}\, dx = \frac{x^2}{2(b \ominus a)} \Bigg|_a^b,$$

which simplifies to

$$\frac{b^2 \ominus a^2}{2(b \ominus a)} = \frac{(b+a)(b \ominus a)}{2(b \ominus a)} = \frac{a+b}{2},$$

which is simply the numerical average of a and b. To find $\mathsf{E}[\cos(X)]$, write

$$\mathsf{E}[\cos(X)] = \int_{-\infty}^{\infty} \cos(x) f(x)\, dx = \int_a^b \cos(x) \frac{1}{b \ominus a}\, dx.$$

Since the antiderivative of cos is sin, we have

$$\mathsf{E}[\cos(X)] \;=\; \frac{\sin(b) \Leftrightarrow \sin(a)}{b \Leftrightarrow a}.$$

In particular, if $b = a + 2\pi k$ for some positive integer k, then $\mathsf{E}[\cos(X)] = 0$.

***Example* 3.6.** Let X be a continuous random variable with standard Gaussian density $f \sim N(0,1)$. Compute $\mathsf{E}[X^n]$ for all $n \geq 1$.

Solution. Write

$$\mathsf{E}[X^n] \;=\; \int_{-\infty}^{\infty} x^n f(x)\, dx,$$

where $f(x) = \exp(\Leftrightarrow x^2/2)/\sqrt{2\pi}$. Since f is an even function of x, the above integrand is odd for n odd. Hence, all the odd moments are zero. For $n \geq 2$, apply integration by parts to

$$\mathsf{E}[X^n] \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{n-1} \cdot x e^{-x^2/2}\, dx$$

to obtain $\mathsf{E}[X^n] = (n \Leftrightarrow 1)\mathsf{E}[X^{n-2}]$, where $\mathsf{E}[X^0] := \mathsf{E}[1] = 1$. When $n = 2$ this yields $\mathsf{E}[X^2] = 1$, and when $n = 4$ this yields $\mathsf{E}[X^4] = 3$. The general result for even n is

$$\mathsf{E}[X^n] \;=\; 1 \cdot 3 \cdots (n \Leftrightarrow 3)(n \Leftrightarrow 1).$$

***Example* 3.7.** Determine $\mathsf{E}[X]$ if X has a Cauchy density with parameter $\lambda = 1$.

Solution. This is a trick question. Recall that as noted following Example 2.14, for signed discrete random variables,

$$\mathsf{E}[X] \;=\; \sum_{i:x_i \geq 0} x_i \wp(X = x_i) + \sum_{i:x_i < 0} x_i \wp(X = x_i),$$

if at least one of the sums is finite. The analogous formula for continuous random variables is

$$\mathsf{E}[X] \;=\; \int_0^{\infty} x f(x)\, dx + \int_{-\infty}^{0} x f(x)\, dx,$$

assuming at least one of the integrals is finite. Otherwise we say that $\mathsf{E}[X]$ is undefined. Write

$$\mathsf{E}[X] \;=\; \int_{-\infty}^{\infty} x f(x)\, dx$$

$$\begin{aligned} &= \int_0^\infty x f(x)\,dx + \int_{-\infty}^0 x f(x)\,dx \\ &= \frac{1}{2\pi}\ln(1+x^2)\,\Big|_0^\infty + \frac{1}{2\pi}\ln(1+x^2)\,\Big|_{-\infty}^0 \\ &= (\infty \Leftrightarrow 0) + (0 \Leftrightarrow \infty) \\ &= \infty \Leftrightarrow \infty \\ &= \text{undefined}. \end{aligned}$$

Moment Generating Functions

The probability generating function was defined in Chapter 2 for discrete random variables taking nonnegative integer values. For more general random variables, we introduce the **moment generating function** (mgf)

$$M_X(s) := \mathsf{E}[e^{sX}].$$

This generalizes the concept of probability generating function because if X is discrete taking only nonnegative integer values, then

$$M_X(s) = \mathsf{E}[e^{sX}] = \mathsf{E}[(e^s)^X] = G_X(e^s).$$

To see why M_X is called the moment generating function, write

$$\begin{aligned} M_X'(s) &= \frac{d}{ds}\mathsf{E}[e^{sX}] \\ &= \mathsf{E}\left[\frac{d}{ds}e^{sX}\right] \\ &= \mathsf{E}[Xe^{sX}]. \end{aligned}$$

Taking $s = 0$, we have

$$M_X'(s)|_{s=0} = \mathsf{E}[X].$$

By differentiating k times and then setting $s = 0$ yields

$$M_X^{(k)}(s)|_{s=0} = \mathsf{E}[X^k]. \tag{3.3}$$

This derivation makes sense only if $M_X(s)$ is finite in a neighborhood of $s = 0$ [4, p. 278].

In the preceding equations, it is convenient to allow s to be complex. This means that we need to define the expectation of complex-valued functions of x. If $g(x) = u(x) + jv(x)$, where u and v are real-valued functions of x, then

$$\mathsf{E}[g(X)] := \mathsf{E}[u(X)] + j\mathsf{E}[v(X)].$$

If X has density f, then

$$\begin{aligned}\mathsf{E}[g(X)] &:= \mathsf{E}[u(X)] + j\mathsf{E}[v(X)] \\ &= \int_{-\infty}^{\infty} u(x)f(x)\,dx + j\int_{-\infty}^{\infty} v(x)f(x)\,dx \\ &= \int_{-\infty}^{\infty} [u(x)+jv(x)]f(x)\,dx \\ &= \int_{-\infty}^{\infty} g(x)f(x)\,dx.\end{aligned}$$

It now follows that if X is a continuous random variable with density f, then

$$M_X(s) \;=\; \mathsf{E}[e^{sX}] \;=\; \int_{-\infty}^{\infty} e^{sx}f(x)\,dx,$$

which is just the **Laplace transform** of f.

***Example* 3.8.** If X is an exponential random variable with parameter $\lambda > 0$, then for $\mathrm{Re}\, s < \lambda$,

$$\begin{aligned}M_X(s) &= \int_0^{\infty} e^{sx}\lambda e^{-\lambda x}\,dx \\ &= \lambda\int_0^{\infty} e^{x(s-\lambda)}\,dx \\ &= \frac{\lambda}{\lambda - s}.\end{aligned}$$

If $M_X(s)$ is finite for all real s in a neighborhood of the origin, say for $-r < s < r$ for some $0 < r \le \infty$, then X has finite moments of all orders, and the following calculation using the power series $e^{\xi} = \sum_{n=0}^{\infty} \xi^n/n!$ is valid for complex s with $|s| < r$ [4, p. 278]:

$$\begin{aligned}\mathsf{E}[e^{sX}] &= \mathsf{E}\left[\sum_{n=0}^{\infty}\frac{(sX)^n}{n!}\right] \\ &= \sum_{n=0}^{\infty}\frac{s^n}{n!}\mathsf{E}[X^n], \quad |s| < r. \end{aligned} \tag{3.4}$$

***Example* 3.9.** For the exponential random variable of the previous example, we can obtain the power series as follows. Recalling the geometric series formula (Problem 7 in Chapter 1), write, this time for $|s| < \lambda$,

$$\frac{\lambda}{\lambda - s} \;=\; \frac{1}{1 - s/\lambda} \;=\; \sum_{n=0}^{\infty}(s/\lambda)^n.$$

Comparing this expression with (3.4) and equating the coefficients of the powers of s, we see by inspection that $\mathsf{E}[X^n] = n!/\lambda^n$. In particular, we have $\mathsf{E}[X] = 1/\lambda$ and $\mathsf{E}[X^2] = 2/\lambda^2$. Since $\mathsf{var}(X) = \mathsf{E}[X^2] \Leftrightarrow (\mathsf{E}[X])^2$, it follows that $\mathsf{var}(X) = 1/\lambda^2$.

***Example* 3.10.** We find the moment generating function of $X \sim N(0,1)$. To begin, write

$$M_X(s) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2}\, dx.$$

We first show that $M_X(s)$ is finite for all real s. Combine the exponents in the above integral and complete the square. This yields

$$M_X(s) \;=\; e^{s^2/2} \int_{-\infty}^{\infty} \frac{e^{-(x-s)^2/2}}{\sqrt{2\pi}}\, dx.$$

The above integrand is a normal density with mean s and unit variance. Since densities integrate to one, $M_X(s) = e^{s^2/2}$ for all real s. Note that since the mean of a real-valued random variable must be real, our argument required that s be real.[5]

We next show that $M_X(s) = e^{s^2/2}$ holds for all complex s. Since the moment generating function is finite for all real s, we can use the power series expansion (3.4) to find M_X. The moments of X were determined in Example 3.6. Recalling that the odd moments are zero,

$$\begin{aligned} M_X(s) &= \sum_{m=0}^{\infty} \frac{s^{2m}}{(2m)!} \mathsf{E}[X^{2m}] \\ &= \sum_{m=0}^{\infty} \frac{s^{2m}}{(2m)!} 1 \cdot 3 \cdots (2m \Leftrightarrow 3)(2m \Leftrightarrow 1) \\ &= \sum_{m=0}^{\infty} \frac{s^{2m}}{2 \cdot 4 \cdots 2m}. \end{aligned}$$

Combining this result with the power series

$$e^{s^2/2} \;=\; \sum_{m=0}^{\infty} \frac{(s^2/2)^m}{m!} \;=\; \sum_{m=0}^{\infty} \frac{s^{2m}}{2 \cdot 4 \cdots 2m}$$

shows that $M_X(s) = e^{s^2/2}$ for all complex s.

Characteristic Functions

In Example 3.8, the moment generating function was guaranteed finite only for $\mathrm{Re}\, s < \lambda$. It is possible to have random variables for which $M_X(s)$ is defined

only for $\mathrm{Re}\, s = 0$; i.e., $M_X(s)$ is only defined for imaginary s. For example, if X is a Cauchy random variable, then it is easy to see that $M_X(s) = \infty$ for all real $s \neq 0$. In order to develop transform methods that always work for *any* random variable X, we introduce the the **characteristic function** of X, defined by

$$\varphi_X(\nu) \;:=\; \mathsf{E}[e^{j\nu X}]. \tag{3.5}$$

Note that $\varphi_X(\nu) = M_X(j\nu)$. Also, since $|e^{j\nu X}| = 1$, $|\varphi_X(\nu)| \le \mathsf{E}[|e^{j\nu X}|] = 1$. Hence, the characteristic function always exists and is bounded in magnitude by one.

If X is a continuous random variable with density f, then

$$\varphi_X(\nu) \;=\; \int_{-\infty}^{\infty} e^{j\nu x} f(x)\, dx,$$

which is just the **Fourier transform** of f. Using the Fourier **inversion formula**,

$$f(x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-j\nu x} \varphi_X(\nu)\, d\nu.$$

***Example* 3.11.** If X is an $N(0,1)$ random variable, then by Example 3.10, $M_X(s) = e^{s^2/2}$. Thus, $\varphi_X(\nu) = M_X(j\nu) = e^{(j\nu)^2/2} = e^{-\nu^2/2}$. In terms of Fourier transforms,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{j\nu x} e^{-x^2/2}\, dx \;=\; e^{-\nu^2/2}.$$

In signal processing terms, the Fourier transform of a **Gaussian pulse** is a Gaussian pulse.

***Example* 3.12.** Let X be a gamma random variable with parameter $p > 0$ (defined in Problem 11). It is shown in Problem 39 that $M_X(s) = 1/(1 \Leftrightarrow s)^p$ for complex s with $|s| < 1$. It follows that $\varphi_X(\nu) = 1/(1 \Leftrightarrow j\nu)^p$ for $|\nu| < 1$. It can be shown[6] that $\varphi_X(\nu) = 1/(1 \Leftrightarrow j\nu)^p$ for all ν.

***Example* 3.13.** As noted above, the characteristic function of an $N(0,1)$ random variable is $e^{-\nu^2/2}$. What is the characteristic function of an $N(m, \sigma^2)$ random variable?

Solution. Let f_0 denote the $N(0,1)$ density. If $X \sim N(m, \sigma^2)$, then $f_X(x) = f_0((x \Leftrightarrow m)/\sigma)/\sigma$. Now write

$$\begin{aligned}
\varphi_X(\nu) &= \mathsf{E}[e^{j\nu X}] \\
&= \int_{-\infty}^{\infty} e^{j\nu x} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} e^{j\nu x} \cdot \frac{1}{\sigma} f_0\left(\frac{x \Leftrightarrow m}{\sigma}\right) dx.
\end{aligned}$$

Now apply the change of variable $y = (x \Leftrightarrow m)/\sigma$, $dy = dx/\sigma$ and obtain

$$\begin{aligned}\varphi_X(\nu) &= \int_{-\infty}^{\infty} e^{j\nu(\sigma y+m)} f_0(y)\, dy \\ &= e^{j\nu m} \int_{-\infty}^{\infty} e^{j(\nu\sigma)y} f_0(y)\, dy \\ &= e^{j\nu m} e^{-(\nu\sigma)^2/2} \\ &= e^{j\nu m - \sigma^2\nu^2/2}.\end{aligned}$$

If X is a discrete *integer-valued* random variable, then

$$\varphi_X(\nu) = \mathsf{E}[e^{j\nu X}] = \sum_n e^{j\nu n} \wp(X = n)$$

is a 2π-periodic **Fourier series**. Given φ_X, the coefficients can be recovered by the formula for Fourier series coefficients,

$$\wp(X = n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-j\nu n} \varphi_X(\nu)\, d\nu.$$

When the moment generating function is not finite in a neighborhood of the origin, the moments of X cannot be obtained from (3.3). However, the moments can sometimes be obtained from the characteristic function. For example, if we differentiate (3.5) with respect to ν, we obtain

$$\varphi_X'(\nu) = \frac{d}{d\nu}\mathsf{E}[e^{j\nu X}] = \mathsf{E}[jXe^{j\nu X}].$$

Taking $\nu = 0$ yields $\varphi_X'(0) = j\mathsf{E}[X]$. The general result is

$$\varphi_X^{(k)}(\nu)|_{\nu=0} = j^k \mathsf{E}[X^k], \quad \text{assuming } \mathsf{E}[|X|^k] < \infty.$$

The assumption that $\mathsf{E}[|X|^k] < \infty$ justifies differentiating inside the expectation [4, pp. 344–345].

3.3. Expectation of Multiple Random Variables

In Chapter 2 we showed that for discrete random variables, expectation is linear and monotone. We also showed that the expectation of a product of independent discrete random variables is the product of the individual expectations. We now derive these properties for arbitrary random variables. Recall that for an arbitrary random variable X, $\mathsf{E}[X] := \lim_{n\to\infty} \mathsf{E}[q_n(X)]$, where $q_n(x)$ is sketched in Figure 3.2, $q_n(x) \to x$, and for each n, $q_n(X)$ is a discrete random variable taking finitely many values.

Linearity of Expectation

To establish linearity, write

$$\begin{aligned} a\mathsf{E}[X] + b\mathsf{E}[Y] &:= a \lim_{n\to\infty} \mathsf{E}[q_n(X)] + b \lim_{n\to\infty} \mathsf{E}[q_n(Y)] \\ &= \lim_{n\to\infty} \mathsf{E}[aq_n(X) + bq_n(Y)] \\ &= \mathsf{E}[\lim_{n\to\infty} aq_n(X) + bq_n(Y)] \\ &= \mathsf{E}[aX + bY]. \end{aligned}$$

From our new definition of expectation, it is clear that if $X \geq 0$, then so is $\mathsf{E}[X]$. Combining this with linearity shows that monotonicity holds for general random variables; i.e., $X \geq Y$ implies $\mathsf{E}[X] \geq \mathsf{E}[Y]$.

Expectations of Products of Functions of Independent Random Variables

Suppose X and Y are independent random variables. For any functions $h(x)$ and $k(y)$ write

$$\begin{aligned} \mathsf{E}[h(X)]\mathsf{E}[k(Y)] &:= \lim_{n\to\infty} \mathsf{E}[q_n(h(X))] \lim_{n\to\infty} \mathsf{E}[q_n(k(Y))] \\ &= \lim_{n\to\infty} \mathsf{E}[q_n(h(X))]\mathsf{E}[q_n(k(Y))] \\ &= \lim_{n\to\infty} \mathsf{E}[q_n(h(X))q_n(k(Y))] \\ &= \mathsf{E}[\lim_{n\to\infty} q_n(h(X))q_n(k(Y))] \\ &= \mathsf{E}[h(X)k(Y)]. \end{aligned}$$

***Example* 3.14.** Let $Z := X + Y$, where X and Y are independent random variables. Show that the characteristic function of Z is the product of the characteristic functions of X and Y.

Solution. The characteristic function of Z is

$$\varphi_Z(\nu) := \mathsf{E}[e^{j\nu Z}] = \mathsf{E}[e^{j\nu(X+Y)}] = \mathsf{E}[e^{j\nu X}e^{j\nu Y}].$$

Now use independence to write

$$\varphi_Z(\nu) = \mathsf{E}[e^{j\nu X}]\mathsf{E}[e^{j\nu Y}] = \varphi_X(\nu)\varphi_Y(\nu).$$

In the preceding example, suppose that X and Y are continuous random variables with densities f_X and f_Y. Then we can inverse Fourier transform the last equation and show that f_Z is the **convolution** of f_X and f_Y. In symbols,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z \Leftrightarrow \tau)\, f_Y(\tau)\, d\tau.$$

***Example* 3.15.** In the preceding example, suppose that X and Y are Cauchy with parameters λ and μ, respectively. Find the density of Z.

Solution. The characteristic of functions of X and Y are, by Problem 44, $\varphi_X(\nu) = e^{-\lambda|\nu|}$ and $\varphi_Y(\nu) = e^{-\mu|\nu|}$. Hence,

$$\varphi_Z(\nu) \;=\; \varphi_X(\nu)\varphi_Y(\nu) \;=\; e^{-\lambda|\nu|}e^{-\mu|\nu|} \;=\; e^{-(\lambda+\mu)|\nu|},$$

which is the characteristic function of a Cauchy random variable with parameter $\lambda + \mu$. In terms of convolution,

$$\frac{(\lambda+\mu)/\pi}{(\lambda+\mu)^2+z^2} \;=\; \int_{-\infty}^{\infty} \frac{\lambda/\pi}{\lambda^2+(z-\tau)^2} \cdot \frac{\mu/\pi}{\mu^2+\tau^2}\, d\tau.$$

3.4. *Probability Bounds

In many applications, it is difficult to compute the probability of an event exactly. However, bounds on the probability can often be obtained in terms of various expectations. For example, the Markov and Chebyshev inequalities were derived in Chapter 2. Below we use Markov's inequality to derive a much stronger result known as the **Chernoff bound**.[‡]

***Example* 3.16** (Using Markov's Inequality). Let X be a Poisson random variable with parameter $\lambda = 1/2$. Use Markov's inequality to bound $\wp(X > 2)$. Compare your bound with the exact result.

Solution. First note that since X takes only integer values, $\wp(X > 2) = \wp(X \geq 3)$. Hence, by Markov's inequality and the fact that $\mathsf{E}[X] = \lambda = 1/2$ from Example 2.13,

$$\wp(X \geq 3) \;\leq\; \frac{\mathsf{E}[X]}{3} \;=\; \frac{1/2}{3} \;=\; 0.167.$$

The exact answer can be obtained by noting that $\wp(X \geq 3) = 1 - \wp(X < 3) = 1 - \wp(X = 0) - \wp(X = 1) - \wp(X = 2)$. For a Poisson($\lambda$) random variable with $\lambda = 1/2$, $\wp(X \geq 3) = 0.0144$. So Markov's inequality gives quite a loose bound.

***Example* 3.17** (Using Chebyshev's Inequality). Let X be a Poisson random variable with parameter $\lambda = 1/2$. Use Chebyshev's inequality to bound $\wp(X > 2)$. Compare your bound with the result of using Markov's inequality in Example 3.16.

[‡] This bound, often attributed to Chernoff (1952) [7], was used earlier by Cramér (1938) [11].

Solution. Since X is nonnegative, we don't have to worry about the absolute value signs. Using Chebyshev's inequality and the fact that $\mathsf{E}[X^2] = \lambda^2 + \lambda = 0.75$ from Example 2.18,

$$\wp(X \geq 3) \;\leq\; \frac{\mathsf{E}[X^2]}{3^2} \;=\; \frac{3/4}{9} \;\approx\; 0.0833.$$

From Example 3.16, the exact probability is 0.0144 and the Markov bound is 0.167.

We now derive the Chernoff bound. Let X be a random variable whose moment generating function $M_X(s)$ is finite for $s \geq 0$. For $s > 0$,

$$\{X \geq a\} \;=\; \{sX \geq sa\} \;=\; \{e^{sX} \geq e^{sa}\}.$$

Using this identity along with Markov's inequality, we have

$$\wp(X \geq a) \;=\; \wp(e^{sX} \geq e^{sa}) \;\leq\; \frac{\mathsf{E}[e^{sX}]}{e^{sa}} \;=\; e^{-sa} M_X(s).$$

Now observe that the inequality holds for all $s > 0$, and the left-hand side does not depend on s. Hence, we can minimize the right-hand side to get as tight a bound as possible. The **Chernoff bound** is given by

$$\wp(X \geq a) \;\leq\; \inf_{s>0} \left[e^{-sa} M_X(s)\right],$$

where the infimum is over all $s > 0$ for which $M_X(s)$ is finite.

***Example* 3.18.** Let X be a Poisson random variable with parameter $\lambda = 1/2$. Bound $\wp(X > 2)$ using the Chernoff bound. Compare your result with the exact probability and with the bound obtained via Chebyshev's inequality in Example 3.17 and with the bound obtained via Markov's inequality in Example 3.16.

Solution. First recall that $M_X(s) = G_X(e^s)$, where $G_X(z) = \exp[\lambda(z \Leftrightarrow 1)]$ was derived in Example 2.19. Hence,

$$e^{-sa} M_X(s) \;=\; e^{-sa} \exp[\lambda(e^s \Leftrightarrow 1)] \;=\; \exp[\lambda(e^s \Leftrightarrow 1) \Leftrightarrow as].$$

The desired Chernoff bound when $a = 3$ is

$$\wp(X \geq 3) \;\leq\; \inf_{s>0} \exp[\lambda(e^s \Leftrightarrow 1) \Leftrightarrow 3s].$$

To evaluate the infimum, we must minimize the exponential. Since exp is an increasing function, it suffices to minimize its argument. Taking the derivative of the argument and setting it equal to zero requires us to solve $\lambda e^s \Leftrightarrow 3 = 0$.

The solution is $s = \ln(3/\lambda)$. Substituting this value of s into $\exp[\lambda(e^s \Leftrightarrow 1) \Leftrightarrow 3s]$ and simplifying the exponent yields

$$\wp(X \geq 3) \;\leq\; \exp[3 \Leftrightarrow \lambda \Leftrightarrow 3\ln(3/\lambda)].$$

Since $\lambda = 1/2$,

$$\wp(X \geq 3) \;\leq\; \exp[2.5 \Leftrightarrow 3\ln 6] \;=\; 0.0564.$$

Recall that from Example 3.16, the exact probability is 0.0144 and Markov's inequality yielded the bound 0.167. From Example 3.17, Chebyshev's inequality yielded the bound 0.0833.

***Example* 3.19.** Let X be a continuous random variable having exponential density with parameter $\lambda = 1$. Compute $\wp(X \geq 7)$ and the corresponding Markov, Chebyshev, and Chernoff bounds.

Solution. The exact probability is $\wp(X \geq 7) = \int_7^\infty e^{-x}\,dx = e^{-7} = 0.00091$. For the Markov and Chebyshev inequalities, recall that from Example 3.9, $\mathsf{E}[X] = 1/\lambda$ and $\mathsf{E}[X^2] = 2/\lambda^2$. For the Chernoff bound, we need $M_X(s) = \lambda/(\lambda \Leftrightarrow s)$ for $s < \lambda$, which was derived in Example 3.8. Armed with these formulas, we find that Markov's inequality yields $\wp(X \geq 7) \leq \mathsf{E}[X]/7 = 1/7 = 0.143$ and Chebyshev's inequality yields $\wp(X \geq 7) \leq \mathsf{E}[X^2]/7^2 = 2/49 = 0.041$. For the Chernoff bound, write

$$\wp(X \geq 7) \;\leq\; \inf_s e^{-7s}/(1 \Leftrightarrow s),$$

where the minimization is over $0 < s < 1$. The derivative of $e^{-7s}/(1 \Leftrightarrow s)$ with respect to s is

$$\frac{e^{-7s}(7s \Leftrightarrow 6)}{(1 \Leftrightarrow s)^2}.$$

Setting this equal to zero requires that $s = 6/7$. Hence, the Chernoff bound is

$$\wp(X \geq 7) \;\leq\; \left.\frac{e^{-7s}}{(1 \Leftrightarrow s)}\right|_{s=6/7} \;=\; 7e^{-6} \;=\; 0.017.$$

For sufficiently large a, the Chernoff bound on $\wp(X \geq a)$ is always smaller than the bound obtained by Chebyshev's inequality, and this is smaller than the one obtained by Markov's inequality. However, for small a, this may not be the case. See Problem 56 for an example.

3.5. Notes

Notes §3.1: Definition and Notation

***Note* 1.** Strictly speaking, it can only be shown that f in (3.1) is nonnegative almost everywhere; that is

$$\int_{\{t\in\mathbb{R}:\, f(t)<0\}} 1\,dt \;=\; 0.$$

For example, f could be negative at a finite or countably infinite set of points.

***Note* 2.** If U and V are $N(0,1)$, then by Problem 41, U^2 and V^2 are each chi-squared with one degree of freedom (defined in Problem 12). By the Remark following Problem 46(c), U^2+V^2 is chi-squared with 2 degrees of freedom, which is the same as $\exp(1/2)$.

Notes §3.2: Expectation of a Single Random Variable

***Note* 3.** Since q_n in Figure 3.2 is defined only for $x \ge 0$, the definition of expectation in the text applies only arbitrary nonnegative random variables. However, for signed random variables,

$$X \;=\; \frac{|X|+X}{2} - \frac{|X|-X}{2},$$

and we define

$$\mathsf{E}[X] \;:=\; \mathsf{E}\left[\frac{|X|+X}{2}\right] - \mathsf{E}\left[\frac{|X|-X}{2}\right],$$

assuming the difference is not of the form $\infty-\infty$. Otherwise, we say $\mathsf{E}[X]$ is undefined.

We also point out that for $x \ge 0$, $q_n(x) \to x$. To see this, fix any $x \ge 0$, and let $n > x$. Then x will lie under one of the steps in Figure 3.2. If x lies under the kth step, then

$$\frac{k-1}{2^n} \;\le\; x \;<\; \frac{k}{2^n},$$

For x in this range, the value of $q_n(x)$ is $(k-1)/2^n$. Hence, $0 \le x - q_n(x) < 1/2^n$.

Another important fact to note is that for each $x \ge 0$, $q_n(x) \le q_{n+1}(x)$. Hence, $q_n(X) \le q_{n+1}(X)$, and so $\mathsf{E}[q_n(X)] \le \mathsf{E}[q_{n+1}(X)]$ as well. In other words, the sequence of real numbers $\mathsf{E}[q_n(X)]$ is nondecreasing. This implies that $\lim_{n\to\infty} \mathsf{E}[q_n(X)]$ exists either as a finite real number or the extended real number ∞ [38, p. 55].

***Note* 4.** In light of the preceding note, we are using Lebesgue's monotone convergence theorem [4, p. 208], which applies to nonnegative functions.

***Note* 5.** If s were complex, we could interpret $\int_{-\infty}^{\infty} e^{-(x-s)^2/2}\,dx$ as a contour integral in the complex plane. By appealing to the Cauchy–Goursat Theorem [9, pp. 115–121], one could then show that this integral is equal to $\int_{-\infty}^{\infty} e^{-t^2/2}\,dt = \sqrt{2\pi}$. Alternatively, one can use a **permanence of form argument** [9, pp. 286–287]. In this approach, one shows that $M_X(s)$ is analytic in some region, in this case the whole complex plane. One then obtains a formula for $M_X(s)$ on a contour in this region, in this case, the contour is the entire real axis. The permanence of form theorem then states that the formula is valid in the entire region.

***Note* 6.** The permanence of form argument mentioned above in Note 5 is the easiest approach. Since $M_X(s)$ is analytic for complex s with $\mathrm{Re}\, s < 1$, and since $M_X(s) = 1/(1 \Leftrightarrow s)^p$ for real s with $s < 1$, $M_X(s) = 1/(1 \Leftrightarrow s)^p$ holds for all complex s with $\mathrm{Re}\, s < 1$. In particular, the formula holds for $s = j\nu$, since in this case, $\mathrm{Re}\, s = 0 < 1$.

3.6. Problems

Problems §3.1: Definition and Notation

1. A certain burglar alarm goes off if its input voltage exceeds five volts at three consecutive sampling times. If the voltage samples are independent and uniformly distributed on $[0, 7]$, find the probability that the alarm sounds.

2. Let X have density

$$f(x) = \begin{cases} 2/x^3, & x > 1, \\ 0, & \text{otherwise}, \end{cases}$$

   The **median** of X is the number t satisfying $\wp(X > t) = 1/2$. Find the median of X.

3. Let X have an $\exp(\lambda)$ density.

   (a) Show that $\wp(X > t) = e^{-\lambda t}$.

   (b) Compute $\wp(X > t + \Delta t | X > t)$. *Hint:* If $A \subset B$, then $A \cap B = A$.

   Remark. Observe that X has a memoryless property similar to that of the geometric$_1(p)$ random variable. See the remark following Problem 4 in Chapter 2.

4. A company produces independent voltage regulators whose outputs are $\exp(\lambda)$ random variables. In a batch of 10 voltage regulators, find the probability that exactly 3 of them produce outputs greater than v volts.

5. Let $X_1, \ldots, X_n$ be i.i.d. $\exp(\lambda)$ random variables.

   (a) Find the probability that $\min(X_1, \ldots, X_n) > 2$.

(b) Find the probability that $\max(X_1, \ldots, X_n) > 2$.

Hint: Example 2.7 may be helpful.

6. A certain computer is equipped with a hard drive whose lifetime (measured in months) is $X \sim \exp(\lambda)$. The lifetime of the monitor (also measured in months) is $Y \sim \exp(\mu)$. Assume the lifetimes are independent.

   (a) Find the probability that the monitor fails during the first 2 months.

   (b) Find the probability that both the hard drive and the monitor fail during the first year.

   (c) Find the probability that either the hard drive or the monitor fails during the first year.

7. A random variable X has the **Weibull** density with parameters $p > 0$ and $\lambda > 0$, denoted by $X \sim \text{Weibull}(p, \lambda)$, if its density is given by $f(x) := \lambda p x^{p-1} e^{-\lambda x^p}$ for $x > 0$, and $f(x) := 0$ for $x \leq 0$.

   (a) Show that this density integrates to one.

   (b) If $X \sim \text{Weibull}(p, \lambda)$, evaluate $\wp(X > t)$ for $t > 0$.

   (c) Let $X_1, \ldots, X_n$ be i.i.d. Weibull(p, λ) random variables. Find the probability that none of them exceeds 3. Find the probability that at least one of them exceeds 3.

   Remark. The Weibull density arises in the study of reliability in Chapter 4. Note that Weibull$(1, \lambda)$ is the same as $\exp(\lambda)$.

*8. The standard normal density $f \sim N(0, 1)$ is given by $f(x) := e^{-x^2/2}/\sqrt{2\pi}$. The following steps provide a mathematical proof that the normal density is indeed "bell-shaped" as shown in Figure 3.1.

   (a) Use the derivative of f to show that f is decreasing for $x > 0$ and increasing for $x < 0$. (It then follows that f has a global maximum at $x = 0$.)

   (b) Show that f is concave for $|x| < 1$ and convex for $|x| > 1$. *Hint:* Show that the second derivative of f is negative for $|x| < 1$ and positive for $|x| > 1$.

   (c) Since $e^z = \sum_{n=0}^{\infty} z^n/n!$, for positive z, $e^z \geq z$. Hence, $e^{+x^2/2} \geq x^2/2$. Use this fact to show that $e^{-x^2/2} \to 0$ as $|x| \to \infty$.

*9. Use the results of the preceding problem to obtain the corresponding three results for the general normal density $f \sim N(m, \sigma^2)$. *Hint:* Let $\varphi(t) := e^{-t^2/2}/\sqrt{2\pi}$ denote the $N(0, 1)$ density, and observe that $f(x) = \varphi((x \Leftrightarrow m)/\sigma)/\sigma$.

*10. As in the preceding problem, let $f \sim N(m, \sigma^2)$. Keeping in mind that $f(x)$ depends on $\sigma > 0$, show that $\lim_{\sigma \to \infty} f(x) = 0$. Show also that for $x \neq m$, $\lim_{\sigma \to 0} f(x) = 0$, whereas for $x = m$, $\lim_{\sigma \to 0} f(x) = \infty$. With $m = 0$, sketch $f(x)$ for $\sigma = 0.5, 1$, and 2.

11. The **gamma** density with parameter $p > 0$ is given by

$$g_p(x) := \begin{cases} \dfrac{x^{p-1}\, e^{-x}}{, (p)}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

where $, (p)$ is the **gamma function**,

$$, (p) := \int_0^\infty x^{p-1}\, e^{-x}\, dx, \quad p > 0.$$

In other words, the gamma *function* is defined exactly so that the gamma *density* integrates to one. Note that the gamma density is a generalization of the exponential since g_1 is the $\exp(1)$ density.

Remark. In MATLAB, $, (p) =$ `gamma(p)`.

(a) Use integration by parts to show that

$$, (p) = (p \Leftrightarrow 1) \cdot , (p \Leftrightarrow 1), \quad p > 1.$$

Since $, (1)$ can be directly evaluated and is equal to one, it follows that

$$, (n) = (n \Leftrightarrow 1)!, \quad n = 1, 2, \ldots.$$

Thus $,$ is sometimes called the **factorial function**.

(b) Show that $, (1/2) = \sqrt{\pi}$ as follows. In the defining integral, use the change of variable $y = \sqrt{2x}$. Write the result in terms of the standard normal density, which integrates to one, in order to obtain the answer by inspection.

(c) Show that

$$, \left(\frac{2n+1}{2}\right) = \frac{(2n \Leftrightarrow 1) \cdots 5 \cdot 3 \cdot 1}{2^n} \sqrt{\pi}, \quad n \geq 1.$$

Remark. The integral definition of $, (p)$ makes sense only for $p > 0$. However, the recursion $, (p) = (p \Leftrightarrow 1), (p \Leftrightarrow 1)$ suggests a simple way to define $,$ for negative, noninteger arguments. For $0 < \varepsilon < 1$, the right-hand side of $, (\varepsilon) = (\varepsilon \Leftrightarrow 1), (\varepsilon \Leftrightarrow 1)$ is undefined. However, we rearrange this equation and make the *definition*,

$$, (\varepsilon \Leftrightarrow 1) := \frac{, (\varepsilon)}{\Leftrightarrow 1 \Leftrightarrow \varepsilon}.$$

Similarly writing , $(\varepsilon \Leftrightarrow 1) = (\varepsilon \Leftrightarrow 2)$, $(\varepsilon \Leftrightarrow 2)$, and so on, leads to

$$, (\varepsilon \Leftrightarrow n) \;=\; \frac{(\Leftrightarrow 1)^n, (\varepsilon)}{(n \Leftrightarrow \varepsilon)\cdots(2 \Leftrightarrow \varepsilon)(1 \Leftrightarrow \varepsilon)}.$$

12. Important generalizations of the gamma density g_p of the preceding problem arise if we include a **scale parameter**. For $\lambda > 0$, put

$$g_{p,\lambda}(x) \;:=\; \lambda\, g_p(\lambda x) \;=\; \lambda \frac{(\lambda x)^{p-1} e^{-\lambda x}}{, (p)}, \quad x > 0.$$

We write $X \sim \text{gamma}(p, \lambda)$ if X has density $g_{p,\lambda}$, which is called the gamma density with parameters p and λ.

(a) Let f be any probability density. For $\lambda > 0$, show that

$$f_\lambda(x) \;:=\; \lambda\, f(\lambda x)$$

is also a probability density.

(b) For $p = m$ a positive integer, $g_{m,\lambda}$ is called the **Erlang** density with parameters m and λ. We write $X \sim \text{Erlang}(m, \lambda)$ if X has density

$$g_{m,\lambda}(x) \;=\; \lambda \frac{(\lambda x)^{m-1}\, e^{-\lambda x}}{(m \Leftrightarrow 1)!}, \quad x > 0.$$

Remark. As you will see in Problem 46(c), the sum of m i.i.d. $\exp(\lambda)$ random variables is $\text{Erlang}(m, \lambda)$.

(c) If $X \sim \text{Erlang}(m, \lambda)$, show that

$$\wp(X > t) \;=\; \sum_{k=0}^{m-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

In other words, if $Y \sim \text{Poisson}(\lambda t)$, then $\wp(X > t) = \wp(Y < m)$.

(d) For $p = k/2$ and $\lambda = 1/2$, $g_{k/2,1/2}$ is called the **chi-squared** density with k degrees of freedom. It is not required that k be an integer. Of course, the chi-squared density with an even number of degrees of freedom, say $k = 2m$, is the same as the $\text{Erlang}(m, 1/2)$ density. Using Problem 11(b), it is also clear that for $k = 1$,

$$g_{1/2,1/2}(x) \;=\; \frac{e^{-x/2}}{\sqrt{2\pi x}}, \quad x > 0.$$

For an odd number of degrees of freedom, say $k = 2m + 1$, where $m \geq 1$, show that

$$g_{\frac{2m+1}{2},\frac{1}{2}}(x) \;=\; \frac{x^{m-1/2} e^{-x/2}}{(2m \Leftrightarrow 1)\cdots 5 \cdot 3 \cdot 1\sqrt{2\pi}}$$

for $x > 0$. *Hint:* Use Problem 11(c).

Remark. As you will see in Problem 41, the chi-squared random variable arises as the square of a normal random variable. In communication systems employing noncoherent receivers, the incoming signal is squared before further processing. Since the thermal noise in these receivers is Gaussian, chi-squared random variables naturally apppear.

*13. The **beta** density with parameters $r > 0$ and $s > 0$ is defined by

$$b_{r,s}(x) := \begin{cases} \dfrac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)} x^{r-1}(1-x)^{s-1}, & 0 < x < 1, \\ 0, & \text{otherwise,} \end{cases}$$

where Γ is the gamma function defined in Problem 11. We note that if $X \sim \text{gamma}(p,\lambda)$ and $Y \sim \text{gamma}(q,\lambda)$ are independent random variables, then $X/(X+Y)$ has the beta density with parameters p and q (Problem 23 in Chapter 5).

(a) Find simplified formulas and sketch the beta density for the following sets of parameter values: (a) $r = 1$, $s = 1$. (b) $r = 2$, $s = 2$. (c) $r = 1/2$, $s = 1$.

(b) Use the following approach to show that the density integrates to one. Write $\Gamma(r)\,\Gamma(s)$ as a product of integrals using different variables of integration, say x and y, respectively. Rewrite this as an iterated integral, with the inner variable of integration being x. In the inner integral, make the change of variable $u = x/(x+y)$. Change the order of integration, and then make the change of variable $v = y/(1-u)$ in the inner integral. Recognize the resulting inner integral, which does not depend on u, as $\Gamma(r+s)$. Thus,

$$\Gamma(r)\,\Gamma(s) = \Gamma(r+s) \int_0^1 u^{r-1}(1-u)^{s-1}\,du, \tag{3.6}$$

showing that indeed, the density integrates to one.

Remark. The above integral, which is a function of r and s, is usually called the **beta function**, and is denoted by $B(r,s)$. Thus,

$$B(r,s) = \frac{\Gamma(r)\,\Gamma(s)}{\Gamma(r+s)}, \tag{3.7}$$

and

$$b_{r,s}(x) = \frac{x^{r-1}(1-x)^{s-1}}{B(r,s)}, \quad 0 < x < 1.$$

*14. Use equation (3.6) in the preceding problem to show that , $(1/2) = \sqrt{\pi}$. *Hint:* Make the change of variable $u = \sin^2\theta$. Then take $r = s = 1/2$.

Remark. In Problem 11, you used the fact that the normal density integrates to one to show that , $(1/2) = \sqrt{\pi}$. Since your derivation there is reversible, it follows that the normal density integrates to one if and only if , $(1/2) = \sqrt{\pi}$. In this problem, you used the fact that the beta density integrates to one to show that , $(1/2) = \sqrt{\pi}$. Thus, you have an alternative derivation of the fact that the normal density integrates to one.

*15. Show that

$$\int_0^{\pi/2} \sin^n\theta\, d\theta \;=\; \frac{\sqrt{\pi}\,,\left(\dfrac{n+1}{2}\right)}{2\,,\left(\dfrac{n+2}{2}\right)}.$$

Hint: Use equation (3.6) in Problem 13 with $r = (n+1)/2$ and $s = 1/2$, and make the substitution $u = \sin^2\theta$.

*16. The beta function $B(r,s)$ is defined as the integral in (3.6) in Problem 13. Show that

$$B(r,s) \;=\; \int_0^{\infty} (1 \Leftrightarrow e^{-\theta})^{r-1} e^{-\theta s}\, d\theta.$$

*17. The **student's *t*** density with ν degrees of freedom is given by

$$f_\nu(x) \;:=\; \frac{\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}, \quad \Leftrightarrow\infty < x < \infty,$$

where B is the beta function. Show that f_ν integrates to one. *Hint:* The change of variable $e^\theta = 1 + x^2/\nu$ may be useful. Also, the result of the preceding problem may be useful.

Remarks. (*i*) Note that $f_1 \sim$ Cauchy(1).

(*ii*) It is shown in Problem 24 in Chapter 5 that if X and Y are independent with $X \sim N(0,1)$ and Y chi-squared with k degrees of freedom, then $X/\sqrt{Y/k}$ has the student's t density with k degrees of freedom, a result of crucial importance in the study of confidence intervals in Chapter 12.

*18. Show that the student's t density $f_\nu(x)$ defined in Problem 17 converges to the standard normal density as $\nu \to \infty$. *Hints:* Recall that

$$\lim_{\nu\to\infty}\left(1+\frac{x}{\nu}\right)^{\nu} \;=\; e^x.$$

Combine (3.7) in Problem 13 with Problem 11 for the cases $\nu = 2n$ and $\nu = 2n+1$ separately. Then appeal to **Wallis's formula** [3, p. 206],

$$\frac{\pi}{2} \;=\; \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\cdots.$$

*19. For p and q positive, let $B(p,q)$ denote the beta function defined by the integral in (3.6) in Problem 13. Show that

$$f_Z(z) \;:=\; \frac{1}{B(p,q)} \cdot \frac{z^{p-1}}{(1+z)^{p+q}}, \quad z > 0,$$

is a valid density (i.e., integrates to one) on $(0,\infty)$. *Hint:* Make the change of variable $t = 1/(1+z)$.

Problems §3.2: Expectation of a Single Random Variable

20. Let X have density $f(x) = 2/x^3$ for $x \geq 1$ and $f(x) = 0$ otherwise. Compute $\mathsf{E}[X]$.

21. Let X have the $\exp(\lambda)$ density, $f_X(x) = \lambda e^{-\lambda x}$, for $x \geq 0$. Use integration by parts to show that $\mathsf{E}[X] = 1/\lambda$.

22. High-Mileage Cars has just begun producing its new Lambda Series, which averages μ miles per gallon. Al's Excellent Autos has a limited supply of n cars on its lot. Actual mileage of the ith car is given by an exponential random variable X_i with $\mathsf{E}[X_i] = \mu$. Assume actual mileages of different cars are independent. Find the probability that at least one car on Al's lot gets less than $\mu/2$ miles per gallon.

23. The **differential entropy** of a continuous random variable X with density f is

$$h(X) \;:=\; \mathsf{E}[\Leftrightarrow \log f(X)] \;=\; \int_{-\infty}^{\infty} f(x) \log \frac{1}{f(x)}\, dx.$$

If $X \sim$ uniform$[0,2]$, find $h(X)$. Repeat for $X \sim$ uniform$[0,\frac{1}{2}]$ and for $X \sim N(m,\sigma^2)$.

24. Let X be a continuous random variable with density f, and suppose that $\mathsf{E}[X] = 0$. If Z is another random variable with density $f_Z(z) := f(z \Leftrightarrow m)$, find $\mathsf{E}[Z]$.

*25. For the random variable X in Problem 20, find $\mathsf{E}[X^2]$.

*26. Let X have the student's t density with ν degrees of freedom, as defined in Problem 17. Show that $\mathsf{E}[|X|^k]$ is finite if and only if $k < \nu$.

27. Let X have moment generating function $M_X(s) = e^{\sigma^2 s^2/2}$. Use formula (3.3) to find $\mathsf{E}[X^4]$.

28. Let $Z \sim N(0,1)$, and put $Y = Z + n$ for some constant n. Show that $\mathsf{E}[Y^4] = n^4 + 6n^2 + 3$.

29. Let $X \sim \text{gamma}(p,1)$ as in Problem 12. Show that

$$\mathsf{E}[X^n] \;=\; \frac{\Gamma(n+p)}{\Gamma(p)} \;=\; p(p+1)(p+2)\cdots(p+[n-1]).$$

30. Let X have the standard **Rayleigh** density, $f(x) := xe^{-x^2/2}$ for $x \ge 0$ and $f(x) := 0$ for $x < 0$.

    (a) Show that $\mathsf{E}[X] = \sqrt{\pi/2}$.

    (b) For $n \ge 2$, show that $\mathsf{E}[X^n] = 2^{n/2}\,\Gamma(1+n/2)$.

    (c) An Internet router has n input links. The flows in the links are independent standard Rayleigh random variables. The router's buffer memory overflows if more than two links have flows greater than β. Find the probability of buffer memory overflow.

31. Let $X \sim \text{Weibull}(p,\lambda)$ as in Problem 7. Show that $\mathsf{E}[X^n] = \Gamma(1+n/p)/\lambda^{n/p}$.

32. A certain nonlinear circuit has random input $X \sim \exp(1)$, and output $Y = X^{1/4}$. Find the second moment of the output.

*33. Let X have the student's t density with ν degrees of freedom, as defined in Problem 17. For n a positive integer less than $\nu/2$, show that

$$\mathsf{E}[X^{2n}] \;=\; \nu^n \frac{\Gamma\left(\frac{2n+1}{2}\right)\Gamma\left(\frac{\nu-2n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}.$$

*34. Recall that the moment generating function of an $N(0,1)$ random variable $e^{s^2/2}$. Use this fact to find the moment generating function of an $N(m,\sigma^2)$ random variable.

35. If $X \sim \text{uniform}(0,1)$, show that $Y = \ln(1/X) \sim \exp(1)$ by finding its moment generating function for $s < 1$.

36. Find a closed-form expression for $M_X(s)$ if $X \sim \text{Laplace}(\lambda)$. Use your result to find $\mathsf{var}(X)$.

*37. Let X be the random variable of Problem 20. For what real values of s is $M_X(s)$ finite? *Hint:* It is not necessary to evaluate $M_X(s)$ to answer the question.

38. Let $M_p(s)$ denote the moment generating function of the gamma density g_p defined in Problem 11. Show that

$$M_p(s) \;=\; \frac{1}{1-s} M_{p-1}(s), \quad p > 1.$$

Remark. Since $g_1(x)$ is the $\exp(1)$ density, and $M_1(s) = 1/(1-s)$ by direct calculation, it now follows that the moment generating function of an $\text{Erlang}(m,1)$ random variable is $1/(1-s)^m$.

*39. To do this problem, use the methods of Example 3.10. Let X have the gamma density g_p given in Problem 11.

(a) For real $s < 1$, show that $M_X(s) = 1/(1 \Leftrightarrow s)^p$.

(b) The moments of X are given in Problem 29. Hence, from (3.4), we have for complex s,

$$M_X(s) = \sum_{n=0}^{\infty} \frac{s^n}{n!} \cdot \frac{, (n+p)}{, (p)}, \quad |s| < 1.$$

For complex s with $|s| < 1$, derive the Taylor series for $1/(1 \Leftrightarrow s)^p$ and show that it is equal to the above series. Thus, $M_X(s) = 1/(1 \Leftrightarrow s)^p$ for all complex s with $|s| < 1$.

40. As shown in the preceding problem, the basic gamma density with parameter p, $g_p(x)$, has moment generating function $1/(1 \Leftrightarrow s)^p$. The more general gamma density defined by $g_{p,\lambda}(x) := \lambda g_p(\lambda x)$ is given in Problem 12.

(a) Find the moment generating function and then the characteristic function of $g_{p,\lambda}(x)$.

(b) Use the answer to (a) to find the moment generating function and the characteristic function of the Erlang density with parameters m and λ, $g_{m,\lambda}(x)$.

(c) Use the answer to (a) to find the moment generating function and the characteristic function of the chi-squared density with k degrees of freedom, $g_{k/2,1/2}(x)$.

41. Let $X \sim N(0,1)$, and put $Y = X^2$. For real values of $s < 1/2$, show that

$$M_Y(s) = \left(\frac{1}{1 \Leftrightarrow 2s}\right)^{1/2}.$$

By Problem 40(c), it follows that Y is chi-squared with one degree of freedom.

42. Let $X \sim N(m,1)$, and put $Y = X^2$. For real values of $s < 1/2$, show that

$$M_Y(s) = \frac{e^{sm^2/(1-2s)}}{\sqrt{1 \Leftrightarrow 2s}}.$$

Remark. For $m \neq 0$, Y is said to be **noncentral chi-squared** with one degree of freedom and **noncentrality parameter** m^2. For $m = 0$, this reduces to the result of the previous problem.

43. Let X have characteristic function $\varphi_X(\nu)$. If $Y := aX + b$ for constants a and b, express the characteristic function of Y in terms of a, b, and φ_X.

44. Apply the Fourier inversion formula to $\varphi_X(\nu) = e^{-\lambda|\nu|}$ to verify that this is the characteristic function of a Cauchy(λ) random variable.

Problems §3.3: Expectation of Multiple Random Variables

45. Let X and Y be independent random variables with moment generating functions $M_X(s)$ and $M_Y(s)$. If $Z := X \Leftrightarrow Y$, show that $M_Z(s) = M_X(s)M_Y(\Leftrightarrow s)$. Show that if both X and Y are exp(λ), then $Z \sim$ Laplace(λ).

46. Let $X_1, \ldots, X_n$ be independent, and put $Y_n := X_1 + \cdots + X_n$.

    (a) If $X_i \sim N(m_i, \sigma_i^2)$, show that $Y_n \sim N(m, \sigma^2)$, and identify m and σ^2. *Hint:* Example 3.13 may be helpful.

    (b) If $X_i \sim$ Cauchy(λ_i), show that $Y_n \sim$ Cauchy(λ), and identify λ. *Hint:* Problem 44 may be helpful.

    (c) If X_i is a gamma random variable with parameters p_i and λ (same λ for all i), show that Y_n is gamma with parameters p and λ, and identify p. *Hint:* The results of Problem 40 may be helpful.

    Remark. Note the following special cases of this result. If all the $p_i = 1$, then the X_i are exponential with parameter λ, and Y_n is Erlang with parameters n and λ. If $p = 1/2$ and $\lambda = 1/2$, then X_i is chi-squared with one degree of freedom, and Y_n is chi-squared with n degrees of freedom.

47. Let $X_1, \ldots, X_r$ be i.i.d. gamma random variables with parameters p and λ. Let $Y = X_1 + \cdots + X_r$. Find $\mathsf{E}[Y^n]$.

48. Packet transmission times on a certain network link are i.i.d. with an exponential density of parameter λ. Suppose n packets are transmitted. Find the density of the time to transmit n packets.

49. The random number generator on a computer produces i.i.d. uniform$(0,1)$ random variables $X_1, \ldots, X_n$. Find the probability density of

$$Y \;=\; \ln\biggl(\prod_{i=1}^{n} \frac{1}{X_i}\biggr).$$

50. Let $X_1, \ldots, X_n$ be i.i.d. Cauchy(λ). Find the density of $Y := \beta_1 X_1 + \cdots + \beta_n X_n$, where the β_i are given positive constants.

51. Three independent pressure sensors produce output voltages U, V, and W, each exp(λ) random variables. The three voltages are summed and fed into an alarm that sounds if the sum is greater than x volts. Find the probability that the alarm sounds.

52. A certain electric power substation has n power lines. The line loads are independent Cauchy(λ) random variables. The substation automatically shuts down if the total load is greater than ℓ. Find the probability of automatic shutdown.

53. The new outpost on Mars extracts water from the surrounding soil. There are 13 extractors. Each extractor produces water with a random efficiency that is uniformly distributed on $[0, 1]$. The outpost operates normally if fewer than 3 extractors produce water with efficiency less than 0.25. If the efficiencies are independent, find the probability that the outpost operates normally.

*54. In this problem we generalize the **noncentral chi-squared** density of Problem 42. To distinguish these new densities from the original chi-squared densities defined in Problem 12, we refer to the original ones as **central** chi-squared densities. The noncentral chi-squared density with k degrees of freedom and **noncentrality parameter** λ^2 is defined by[§]

$$c_{k,\lambda^2}(x) \;:=\; \sum_{n=0}^{\infty} \frac{(\lambda^2/2)^n e^{-\lambda^2/2}}{n!}\, c_{2n+k}(x), \quad x > 0,$$

where c_{2n+k} denotes the central chi-squared density with $2n + k$ degrees of freedom.

(a) Show that $\int_0^\infty c_{k,\lambda^2}(x)\,dx = 1$.

(b) If X is a noncentral chi-squared random variable with k degrees of freedom and noncentrality parameter λ^2, show that X has moment generating function

$$M_{k,\lambda^2}(s) \;=\; \frac{\exp[s\lambda^2/(1 \Leftrightarrow 2s)]}{(1 \Leftrightarrow 2s)^{k/2}}.$$

Hint: Problem 40 may be helpful.

Remark. When $k = 1$, this agrees with Problem 42.

(c) Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim c_{k_i,\lambda_i^2}$. Show that $Y := X_1 + \cdots + X_n$ has the c_{k,λ^2} density, and identitify k and λ^2.

Remark. By part (b), if each $k_i = 1$, we could assume that each X_i is the *square* of an $N(\lambda_i, 1)$ random variable.

(d) Show that

$$\frac{e^{-(x+\lambda^2)/2}}{\sqrt{2\pi x}} \cdot \frac{e^{\lambda\sqrt{x}} + e^{-\lambda\sqrt{x}}}{2} \;=\; c_{1,\lambda^2}(x).$$

(Note that if $\lambda = 0$, the left-hand side reduces to the central chi-squared density with one degree of freedom.) *Hint:* Use the power series $e^\xi = \sum_{n=0}^\infty \xi^n/n!$ for the two exponentials involving $\sqrt{x}$.

[§]A closed-form expression is derived in Problem 17 of Chapter 4.

Problems §3.4: Probability Bounds

55. Let X be the random variable of Problem 20. For $a \geq 1$, compare $\wp(X \geq a)$ and the bound obtained via Markov's inequality.

*56. Let X be an exponential random variable with parameter $\lambda = 1$. Use Markov's inequality, Chebyshev's inequality, and the Chernoff bound to obtain bounds on $\wp(X \geq a)$ as a function of a .

(a) For what values of a does Markov's inequality give the best bound?

(b) For what values of a does Chebyshev's inequality give the best bound?

(c) For what values of a does the Chernoff bound work best?

CHAPTER 4

Analyzing Systems with Random Inputs

In this chapter we consider systems with random inputs, and we try to compute probabilities involving the system outputs. The key tool we use to solve these problems is the **cumulative distribution function** (cdf). If X is a real-valued random variable, its cdf is defined by

$$F_X(x) \;:=\; \wp(X \le x).$$

Cumulative distribution functions of continuous random variables are introduced in Section 4.1. The emphasis is on the fact that if $Y = g(X)$, where X is a continuous random variable and g is a fairly simple function, then the density of Y can be found by first determining the cdf of Y via simple algebraic manipulations and then differentiating.

The material in Section 4.2 is a brief diversion into reliability theory. The topic is placed here because it makes simple use of the cdf. With the exception of the formula

$$\mathsf{E}[T] \;=\; \int_0^\infty \wp(T > t)\, dt$$

for nonnegative random variables, which is derived at the beginning of Section 4.2, the remaining material on reliability is not used in the rest of the book.

In trying to compute probabilities involving $Y = g(X)$, the method of Section 4.1 only works for very simple functions g. To handle more complicated functions, we need more background on cdfs.* Section 4.3 provides a brief introduction to cdfs of discrete random variables. This section is not of great interest in itself, but serves as preparation for Section 4.4 on mixed random variables. Mixed random variables frequently appear in the form $Y = g(X)$ when X is continuous, but g has "flat spots." More precisely, there is an interval where $g(x)$ is constant. For example, most amplifiers have a linear region, say $\Leftrightarrow v \le x \le v$, where $g(x) = \alpha x$. But if $x > v$, $g(x) = \alpha v$, and if $x < \Leftrightarrow v$, $g(x) = \Leftrightarrow \alpha v$. In this case, if a continuous random variable is applied to such a device, the output will be a mixed random variable, which can be thought of has a random variable whose "generalized density" contains Dirac impulses. The problem of finding the cdf and generalized density of $Y = g(X)$ is studied in Section 4.5.

At this point, having seen several generalized densities and their corresponding cdfs, Section 4.6 summarizes and derives the general properties that characterize arbitrary cdfs.

Section 4.7 introduces the **central limit theorem**. Although we have seen many examples for which we can explicitly write down probabilities involving a

*The material in Sections 4.3–4.5 is not used in the rest of the book, though it does provide examples of cdfs with jump discontinuities.

sum of i.i.d. random variables, in general, the problem is very hard. The central limit theorem provides an approximation for computing probabilities involving the sum of i.i.d. random variables.

4.1. Continuous Random Variables

If X is a continuous random variable with density f, then

$$F_X(x) \;=\; \wp(X \le x) \;=\; \int_{-\infty}^{x} f(t)\,dt.$$

***Example* 4.1.** Find the cdf of a Cauchy random variable X with parameter $\lambda = 1$.

Solution. Write

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \frac{1/\pi}{1+t^2}\,dt \\ &= \frac{1}{\pi}\tan^{-1}(t)\,\Big|_{-\infty}^{x} \\ &= \frac{1}{\pi}\left(\tan^{-1}(x) \Leftrightarrow \frac{\Leftrightarrow\pi}{2}\right) \\ &= \frac{1}{\pi}\tan^{-1}(x) + \frac{1}{2}. \end{aligned}$$

A graph of F_X is shown in Figure 4.1.

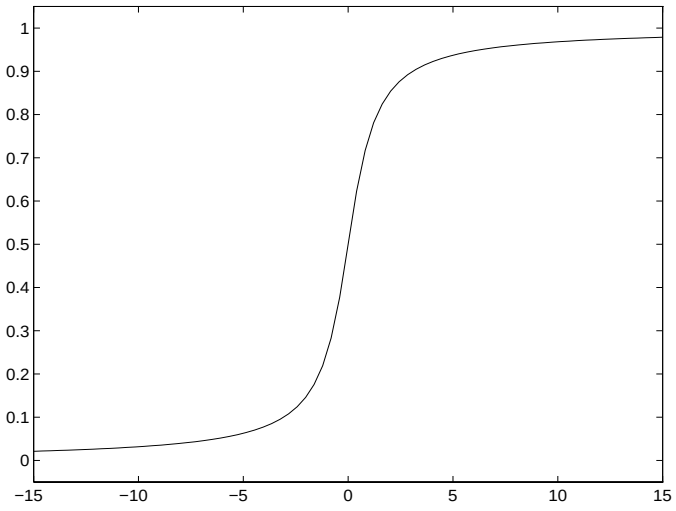

Figure 4.1. Cumulative distribution function of a Cauchy random variable.

***Example* 4.2.** Find the cdf of a uniform$[a,b]$ random variable X.

Solution. Since $f(t) = 0$ for $t < a$ and $t > b$, we see that $F_X(x) = \int_{-\infty}^{x} f(t)\,dt$ is equal to 0 for $x < a$, and is equal to $\int_{-\infty}^{\infty} f(t)\,dt = 1$ for $x > b$. For $a \le x \le b$, we have

$$F_X(x) \;=\; \int_{-\infty}^{x} f(t)\,dt \;=\; \int_{a}^{x} \frac{1}{b \hat{-} a}\,dt \;=\; \frac{x \hat{-} a}{b \hat{-} a}.$$

Hence, for $a \le x \le b$, $F_X(x)$ is an **affine**† function of x. A graph of F_X when $X \sim$ uniform$[0,1]$ is shown if Figure 4.2.

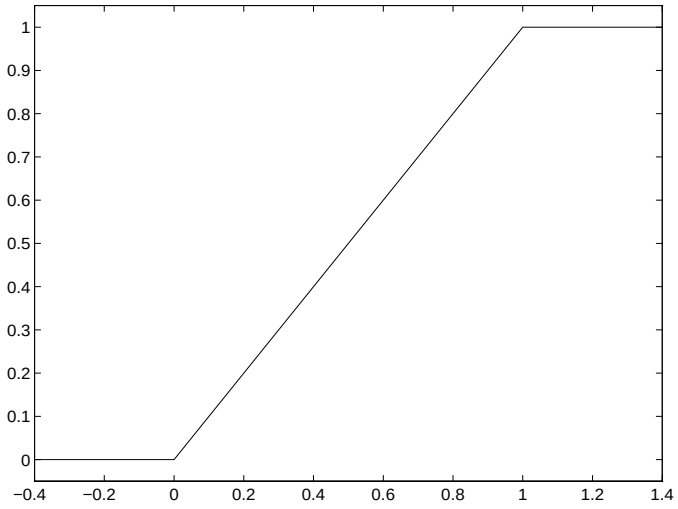

Figure 4.2. Cumulative distribution function of a uniform random variable.

We now consider the cdf of a Gaussian random variable. If $X \sim N(m, \sigma^2)$, then

$$F_X(x) \;=\; \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\hat{-}\frac{1}{2}\left(\frac{t \hat{-} m}{\sigma}\right)^2\right] dt.$$

Unfortunately, there is no closed-form expression for this integral. However, it can be computed numerically, and there are many subroutines available for doing it. For example, in MATLAB, the above integral can be computed with `normcdf(x,m,sigma)`. If you do not have access to such a routine, you may have software available for computing some related functions that can be used to calculate the normal cdf. To see how, we need the following analysis. Making

†A function is affine if it is equal to a linear function plus a constant.

the change of variable $\theta = (t \Leftrightarrow m)/\sigma$ yields

$$F_X(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-m)/\sigma} e^{-\theta^2/2}\, d\theta.$$

In other words, if we put

$$\Phi(y) \;:=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\theta^2/2}\, d\theta, \tag{4.1}$$

then

$$F_X(x) \;=\; \Phi\left(\frac{x \Leftrightarrow m}{\sigma}\right).$$

Note that Φ is the cdf of a standard normal random variable (a graph is shown in Figure 4.3). Thus, the cdf of an $N(m, \sigma^2)$ random variable can always be expressed in terms of the cdf of a standard $N(0, 1)$ random variable. If your computer has a routine that can compute Φ, then you can compute an arbitrary normal cdf.

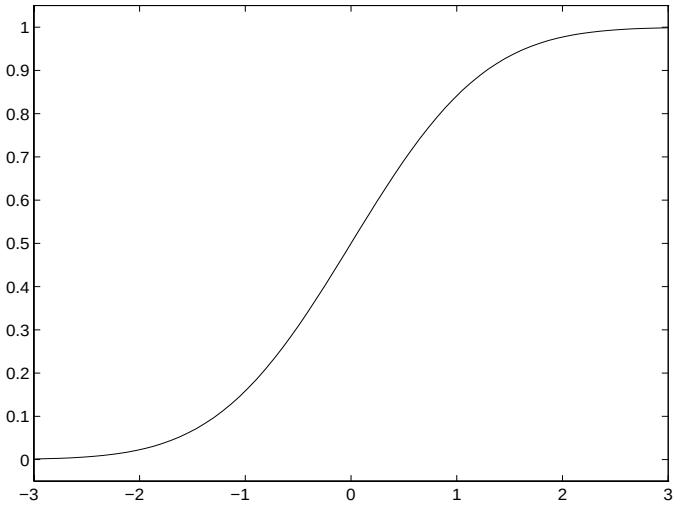

Figure 4.3. Cumulative distribution function of a standard normal random variable.

For continuous random variables, the density can be recovered from the cdf by differentiation. Since

$$F_X(x) \;=\; \int_{-\infty}^{x} f(t)\, dt,$$

differentiation yields

$$F'(x) \;=\; f(x).$$

The observation that the density of a continuous random variable can be recovered from its cdf is of tremendous importance, as the following examples illustrate.

***Example* 4.3.** Consider an electrical circuit whose random input voltage X is first amplified by a gain $\mu > 0$ and then added to a constant offset voltage β. Then the output voltage is $Y = \mu X + \beta$. If the input voltage is a continuous random variable X, find the density of the output voltage Y.

Solution. Although the question asks for the density of Y, it is more advantageous to find the cdf first and then differentiate to obtain the density. Write

$$\begin{aligned} F_Y(y) &= \wp(Y \le y) \\ &= \wp(\mu X + \beta \le y) \\ &= \wp(X \le (y \Leftrightarrow \beta)/\mu), \quad \text{since } \mu > 0, \\ &= F_X((y \Leftrightarrow \beta)/\mu). \end{aligned}$$

If X has density f_X, then

$$f_Y(y) = \frac{d}{dy} F_X\left(\frac{y \Leftrightarrow \beta}{\mu}\right) = \frac{1}{\mu} F_X'\left(\frac{y \Leftrightarrow \beta}{\mu}\right) = \frac{1}{\mu} f_X\left(\frac{y \Leftrightarrow \beta}{\mu}\right).$$

***Example* 4.4.** Amplitude modulation in certain communication systems can be accomplished using various nonlinear devices such as a semiconductor diode. Suppose we model the nonlinear device by the function $Y = X^2$. If the input X is a continuous random variable, find the density of the output $Y = X^2$.

Solution. Although the question asks for the density of Y, it is more advantageous to find the cdf first and then differentiate to obtain the density. To begin, note that since $Y = X^2$ is nonnegative, for $y < 0$, $F_Y(y) = \wp(Y \le y) = 0$. For nonnegative y, write

$$\begin{aligned} F_Y(y) &= \wp(Y \le y) \\ &= \wp(X^2 \le y) \\ &= \wp(\Leftrightarrow\sqrt{y} \le X \le \sqrt{y}\,) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} f_X(t)\, dt. \end{aligned}$$

The density is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} \int_{-\sqrt{y}}^{\sqrt{y}} f_X(t)\, dt \\ &= \frac{1}{2\sqrt{y}} \big[f_X(\sqrt{y}\,) + f_X(\Leftrightarrow\sqrt{y}\,) \big], \quad y > 0. \end{aligned}$$

Since $\wp(Y \le y) = 0$ for $y < 0$, $f_Y(y) = 0$ for $y < 0$.

When the diode input voltage X of the preceding example is $N(0,1)$, it turns out that Y is chi-squared with one degree of freedom (Problem 7). If X is $N(m,1)$ with $m \neq 0$, then Y is noncentral chi-squared with one degree of freedom (Problem 8). These results are frequently used in the analysis of digital communication systems.

The two preceding examples illustrate the problem of finding the density of $Y = g(X)$ when X is a continuous random variable. The next example illustrates the problem of finding the density of $Z = g(X,Y)$ when X is discrete and Y is continuous.

***Example* 4.5** (Signal in Additive Noise). Let X and Y be independent random variables, with X being discrete with pmf p_X and Y being continuous with density f_Y. Put $Z := X + Y$ and find the density of Z.

Solution. Although the question asks for the density of Z, it is more advantageous to find the cdf first and then differentiate to obtain the density. This time we use the law of total probability, substitution, and independence. Write

$$\begin{aligned}
F_Z(z) &= \wp(Z \le z) \\
&= \sum_i \wp(Z \le z | X = x_i)\wp(X = x_i) \\
&= \sum_i \wp(X + Y \le z | X = x_i)\wp(X = x_i) \\
&= \sum_i \wp(x_i + Y \le z | X = x_i)\wp(X = x_i) \\
&= \sum_i \wp(Y \le z \Leftrightarrow x_i | X = x_i)\wp(X = x_i) \\
&= \sum_i \wp(Y \le z \Leftrightarrow x_i)\wp(X = x_i) \\
&= \sum_i F_Y(z \Leftrightarrow x_i)\, p_X(x_i).
\end{aligned}$$

Differentiating this expression yields

$$f_Z(z) = \sum_i f_Y(z \Leftrightarrow x_i)\, p_X(x_i).$$

We should also note that $F_{Z|X}(z|x_i) := \wp(Z \le z | X = x_i)$ is called the **conditional cdf** of Z given X.

* The Normal CDF and the Error Function

If your computer does not have a routine that computes the normal cdf, it may have a routine to compute a related function called the **error function**, which is defined by

$$\operatorname{erf}(z) \;:=\; \frac{2}{\sqrt{\pi}} \int_0^z e^{-\xi^2}\, d\xi.$$

Note that $\operatorname{erf}(z)$ is negative for $z < 0$. In fact, erf is an odd function. To express Φ in terms of erf, write

$$\begin{aligned}
\Phi(y) &= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^0 e^{-\theta^2/2}\, d\theta + \int_0^y e^{-\theta^2/2}\, d\theta \right] \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^y e^{-\theta^2/2}\, d\theta, \quad \text{by Example 3.4,} \\
&= \frac{1}{2} + \frac{1}{2} \cdot \frac{2}{\sqrt{\pi}} \int_0^{y/\sqrt{2}} e^{-\xi^2}\, d\xi,
\end{aligned}$$

where the last step follows by making the change of variable $\xi = \theta/\sqrt{2}$. We now see that

$$\Phi(y) \;=\; \tfrac{1}{2}\left[1 + \operatorname{erf}\left(y/\sqrt{2}\,\right)\right].$$

From the derivation, it is easy to see that this holds for all y, both positive and negative. The MATLAB command for $\operatorname{erf}(z)$ is `erf(z)`.

In many applications, instead of $\Phi(y)$, we need

$$Q(y) \;:=\; 1 \Leftrightarrow \Phi(y).$$

Using the **complementary error function**,

$$\operatorname{erfc}(z) \;:=\; 1 \Leftrightarrow \operatorname{erf}(z) \;=\; \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-\xi^2}\, d\xi,$$

it is easy to see that

$$Q(y) \;=\; \tfrac{1}{2} \operatorname{erfc}\left(y/\sqrt{2}\,\right).$$

The MATLAB command for $\operatorname{erfc}(z)$ is `erfc(z)`.

4.2. Reliability

Let T be the lifetime of a device or system. The **reliability function** of the device or system is defined by

$$R(t) \;:=\; \wp(T > t) \;=\; 1 \Leftrightarrow F_T(t).$$

The reliability at time t is the probability that the lifetime is greater than t.

The **mean time to failure** (MTTF) is simply the expected lifetime, $\mathsf{E}[T]$. Since lifetimes are nonnegative random variables, we claim that

$$\mathsf{E}[T] \;=\; \int_0^\infty \wp(T > t)\, dt. \tag{4.2}$$

It then follows that

$$\mathsf{E}[T] \;=\; \int_0^\infty R(t)\,dt;$$

i.e., the MTTF is the integral of the reliability. To derive (4.2), write

$$\begin{aligned}\int_0^\infty \wp(T>t)\,dt &= \int_0^\infty \mathsf{E}[I_{(t,\infty)}(T)]\,dt \\ &= \mathsf{E}\bigg[\int_0^\infty I_{(t,\infty)}(T)\,dt\bigg] \\ &= \mathsf{E}\bigg[\int_0^\infty I_{(-\infty,T)}(t)\,dt\bigg].\end{aligned}$$

To evaluate the integral, observe that since T is nonnegative, the intersection of $[0,\infty)$ and $(\Leftrightarrow\infty, T)$ is $[0,T)$. Hence,

$$\int_0^\infty \wp(T>t)\,dt \;=\; \mathsf{E}\bigg[\int_0^T dt\bigg] \;=\; \mathsf{E}[T].$$

The **failure rate** of a device or system with lifetime T is

$$r(t) \;:=\; \lim_{\Delta t\downarrow 0} \frac{\wp(T\le t+\Delta t|T>t)}{\Delta t}.$$

This can be rewritten as

$$\wp(T\le t+\Delta t|T>t) \;\approx\; r(t)\Delta t.$$

In other words, given that the device or system has operated for more than t units of time, the conditional probability of failure before time $t+\Delta t$ is approximately $r(t)\Delta t$. Intuitively, the form of a failure rate function should be as shown in Figure 4.4. For small values of t, $r(t)$ is relatively large when

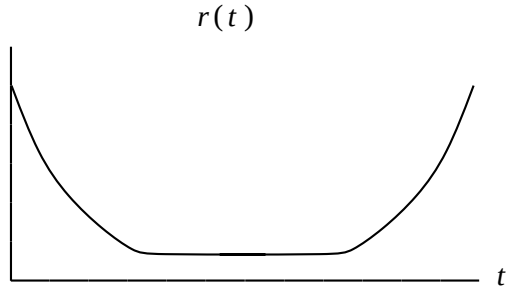

Figure 4.4. Typical form of a failure rate function.

pre-existing defects are likely to appear. Then for intermediate values of t, $r(t)$ is flat indicating a constant failure rate. For large t, as the device gets older, $r(t)$ increases indicating that failure is more likely.

To say more about the failure rate, write

$$\begin{aligned}\wp(T \le t+\Delta t | T > t) &= \frac{\wp(\{T \le t+\Delta t\} \cap \{T > t\})}{\wp(T > t)} \\ &= \frac{\wp(t < T \le t+\Delta t)}{\wp(T > t)} \\ &= \frac{F_T(t+\Delta t) \Leftrightarrow F_T(t)}{R(t)}.\end{aligned}$$

Since $F_T(t) = 1 \Leftrightarrow R(t)$, we can rewrite this as

$$\wp(T \le t+\Delta t | T > t) = \Leftrightarrow \frac{R(t+\Delta t) \Leftrightarrow R(t)}{R(t)}.$$

Dividing both sides by Δt and letting $\Delta t \downarrow 0$ yields the differential equation

$$r(t) = \Leftrightarrow \frac{R'(t)}{R(t)}.$$

Now suppose T is a continuous random variable with density f_T. Then

$$R(t) = \wp(T > t) = \int_t^\infty f_T(\theta)\, d\theta,$$

and

$$R'(t) = \Leftrightarrow f_T(t).$$

We can now write

$$r(t) = \Leftrightarrow \frac{R'(t)}{R(t)} = \frac{f_T(t)}{\int_t^\infty f_T(\theta)\, d\theta}.$$

In this case, the failure rate $r(t)$ is completely determined by the density $f_T(t)$. The converse is also true; i.e., given the failure rate $r(t)$, we can recover the density $f_T(t)$. To see this, rewrite the above differential equation as

$$\Leftrightarrow r(t) = \frac{R'(t)}{R(t)} = \frac{d}{dt} \ln R(t).$$

Integrating the left and right-hand formulas from zero to t yields

$$\Leftrightarrow \int_0^t r(\tau)\, d\tau = \ln R(t) \Leftrightarrow \ln R(0).$$

Then

$$e^{-\int_0^t r(\tau)\, d\tau} = \frac{R(t)}{R(0)} = R(t),$$

where we have used the fact that for a nonnegative, continuous random variable, $R(0) = \wp(T > 0) = \wp(T \ge 0) = 1$. Since $R'(t) = \Leftrightarrow f_T(t)$,

$$f_T(t) = r(t) e^{-\int_0^t r(\tau)\, d\tau}.$$

For example, if the failure rate is constant, $r(t) = \lambda$, then $f_T(t) = \lambda e^{-\lambda t}$, and we see that T has an exponential density with parameter λ.

4.3. Cdfs for Discrete Random Variables

In this section, we show that for discrete random variables, the pmf can be recovered from the cdf. If X is a discrete random variable, then

$$F_X(x) \;=\; \mathcal{P}(X \le x) \;=\; \sum_i I_{(-\infty,x]}(x_i)\, p_X(x_i) \;=\; \sum_{i:x_i \le x} p_X(x_i).$$

***Example* 4.6.** Let X be a discrete random variable with $p_X(0) = 0.1$, $p_X(1) = 0.4$, $p_X(2.5) = 0.2$, and $p_X(3) = 0.3$. Find the cdf $F_X(x)$ for all x.

Solution. Put $x_0 = 0$, $x_1 = 1$, $x_2 = 2.5$, and $x_3 = 3$. Let us first evaluate $F_X(0.5)$. Observe that

$$F_X(0.5) \;=\; \sum_{i:x_i \le 0.5} p_X(x_i) \;=\; p_X(0),$$

since only $x_0 = 0$ is less than or equal to 0.5. Similarly, for any $0 \le x < 1$, $F_X(x) = p_X(0) = 0.1$.

Next observe that

$$F_X(2.1) \;=\; \sum_{i:x_i \le 2.1} p_X(x_i) \;=\; p_X(0) + p_X(1),$$

since only $x_0 = 0$ and $x_1 = 1$ are less than or equal to 2.1. Similarly, for any $1 \le x < 2.5$, $F_X(x) = 0.1 + 0.4 = 0.5$.

The same kind of argument shows that for $2.5 \le x < 3$, $F_X(x) = 0.1 + 0.4 + 0.2 = 0.7$.

Since all the x_i are less than or equal to 3, for $x \ge 3$, $F_X(x) = 0.1 + 0.4 + 0.2 + 0.3 = 1.0$.

Finally, since none of the x_i is less than zero, for $x < 0$, $F_X(x) = 0$. The graph of F_X is given in Figure 4.5.

We now show that the arguments used in the preceding example hold for the cdf of any discrete random variable X taking distinct values x_i. Fix two adjacent points, say $x_{j-1} < x_j$ as shown in Figure 4.6. For $x_{j-1} \le x < x_j$, observe that

$$F_X(x) \;=\; \sum_{i:x_i \le x} p_X(x_i) \;=\; \sum_{i:x_i \le x_{j-1}} p_X(x_i) \;=\; F_X(x_{j-1}).$$

In other words, the cdf is constant on the interval $[x_{j-1}, x_j)$, with the constant equal to the value of the cdf at the left-hand endpoint, $F_X(x_{j-1})$. Next, since

$$F_X(x_j) \;=\; \sum_{i:x_i \le x_j} p_X(x_i) \;=\; \left[\sum_{i:x_i \le x_{j-1}} p_X(x_i)\right] + p_X(x_j),$$

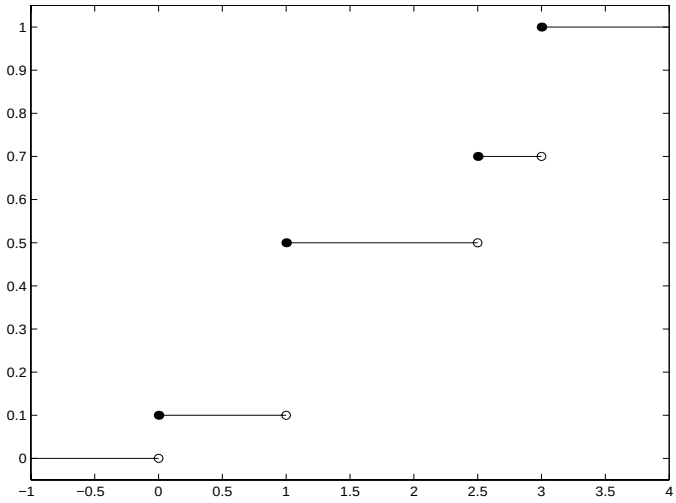

Figure 4.5. Cumulative distribution function of a discrete random variable.

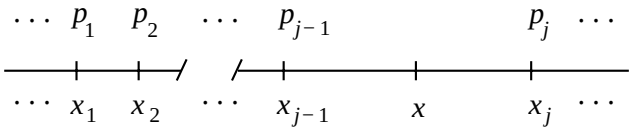

Figure 4.6. Analysis for the cdf of a discrete random variable.

we have

$$F_X(x_j) \;=\; F_X(x_{j-1}) + p_X(x_j),$$

or

$$F_X(x_j) \Leftrightarrow F_X(x_{j-1}) \;=\; p_X(x_j).$$

As noted above, for $x_{j-1} \le x < x_j$, $F_X(x) = F_X(x_{j-1})$. Hence, $F_X(x_j\Leftrightarrow) := \lim_{x\uparrow x_j} F_X(x) = F_X(x_{j-1})$. It then follows that

$$F_X(x_j) \Leftrightarrow F_X(x_j\Leftrightarrow) \;= F_X(x_j) \Leftrightarrow F_X(x_{j-1}) \;=\; p_X(x_j).$$

In other words, the size of the jump in the cdf at x_j is $p_X(x_j)$, and the cdf is constant between jumps.

4.4. Mixed Random Variables

We say that X is a **mixed random variable** if $\wp(X \in B)$ has the form

$$\wp(X \in B) \;=\; \int_B \tilde{f}(t)\,dt + \sum_i I_B(x_i)\tilde{p}_i \tag{4.3}$$

for some nonnegative function $\tilde{f}$, some nonnegative sequence $\tilde{p}_i$, and distinct points x_i such that $\int_{-\infty}^{\infty} \tilde{f}(t)\,dt + \sum_i \tilde{p}_i = 1$. Note that the mixed random variables include the continuous and discrete real-valued random variables as special cases.

***Example* 4.7.** Let X be a mixed random variable with

$$\wp(X \in B) \;:=\; \frac{1}{4}\int_B e^{-|t|}\,dt + \frac{1}{3}I_B(0) + \frac{1}{6}I_B(7). \tag{4.4}$$

Compute $\wp(X < 0)$, $\wp(X \le 0)$, $\wp(X > 2)$, and $\wp(X \ge 2)$.

Solution. We need to compute $\wp(X \in B)$ for $B = (\Leftrightarrow\infty, 0)$, $(\Leftrightarrow\infty, 0]$, $(2, \infty)$, and $[2, \infty)$, respectively. Since the set $B = (\Leftrightarrow\infty, 0)$ does not contain 0 or 7, the two indicator functions in the definition of $\wp(X \in B)$ are zero in this case. Hence,

$$\wp(X < 0) \;=\; \frac{1}{4}\int_{-\infty}^{0} e^{-|t|}\,dt \;=\; \frac{1}{4}\int_{-\infty}^{0} e^{t}\,dt \;=\; \frac{1}{4}.$$

Next, the set $B = (\Leftrightarrow\infty, 0]$ contains 0 but not 7. Thus,

$$\wp(X \le 0) \;=\; \frac{1}{4}\int_{-\infty}^{0} e^{-|t|}\,dt + \frac{1}{3} \;=\; \frac{1}{4} + \frac{1}{3} \;=\; \frac{7}{12}.$$

For the last two probabilities, we need to consider $(2, \infty)$ and $[2, \infty)$. Both intervals contain 7 but not 0. Hence, the probabilities are the same, and both equal

$$\frac{1}{4}\int_2^{\infty} e^{-t}\,dt + \frac{1}{6} \;=\; \frac{e^{-2}}{4} + \frac{1}{6}.$$

***Example* 4.8.** With X as in the last example, compute $\wp(X = 0)$ and $\wp(X = 7)$. Also, use your results to verify that $\wp(X = 0) + \wp(X < 0) = \wp(X \le 0)$.

Solution. To compute $\wp(X = 0) = \wp(X \in \{0\})$, take $B = \{0\}$ in (4.4). Then

$$\wp(X = 0) \;=\; \frac{1}{4}\int_{\{0\}} e^{-|t|}\,dt + \frac{1}{3}I_{\{0\}}(0) + \frac{1}{6}I_{\{0\}}(7).$$

The integral of an ordinary function over a set consisting of a single point is zero, and since $7 \notin \{0\}$, the last term is zero also. Hence, $\wp(X = 0) = 1/3$. Similarly, $\wp(X = 7) = 1/6$. From the previous example, $\wp(X < 0) = 1/4$. Hence, $\wp(X = 0) + \wp(X < 0) = 1/3 + 1/4 = 7/12$, which is indeed the value of $\wp(X \le 0)$ computed earlier.

In order to perform calculations with mixed random variables more easily, it is sometimes convenient to generalize the notion of density to permit impulses. Recall that the **unit impulse** or **Dirac delta function**, denoted by δ, is defined by the two properties

$$\delta(t) \;=\; 0 \text{ for } t \neq 0 \quad \text{and,} \quad \int_{-\infty}^{\infty} \delta(t)\,dt \;=\; 1.$$

Putting

$$f(t) \;:=\; \tilde{f}(t) + \sum_i \tilde{p}_i \delta(t \Leftrightarrow x_i), \tag{4.5}$$

it is easy to verify that

$$\int_B f(t)\,dt = \wp(X \in B) \quad \text{and} \quad \int_{-\infty}^{\infty} g(t) f(t)\,dt = \mathsf{E}[g(X)].$$

We call f in (4.5) a **generalized probability density function**. If $\tilde{f}(t) = 0$ for all t, we say that f is **purely impulsive**, and if $\tilde{p}_i = 0$ for all i, we say f is **nonimpulsive**.

***Example* 4.9.** The preceding examples can be worked using delta functions. In these examples, the generalized density is

$$f(t) \;=\; \frac{1}{4}e^{-|t|} + \frac{1}{3}\delta(t) + \frac{1}{6}\delta(t \Leftrightarrow 7),$$

and

$$\wp(X \in B) \;=\; \int_B f(t)\,dt.$$

Since the impulses are located at 0 and at 7, they will contribute to the integral if and only if 0 and/or 7 lie in the set B. For example, in computing

$$\wp(0 < X \leq 7) \;=\; \int_{0+}^{7} f(t)\,dt,$$

the impulse at the origin makes no contribution, but the impulse at 7 does. Thus,

$$\begin{aligned}
\wp(0 < X \leq 7) \;&=\; \int_{0+}^{7} \left[\frac{1}{4}e^{-|t|} + \frac{1}{3}\delta(t) + \frac{1}{6}\delta(t \Leftrightarrow 7)\right] dt \\
&=\; \frac{1}{4}\int_0^7 e^{-t}\,dt + \frac{1}{6} \\
&=\; \frac{1 \Leftrightarrow e^{-7}}{4} + \frac{1}{6} \;=\; \frac{5}{12} \Leftrightarrow \frac{e^{-7}}{4}.
\end{aligned}$$

Similarly, in computing $\wp(X = 0) = \wp(X \in \{0\})$, only the impulse at zero makes a contribution. Thus,

$$\wp(X = 0) \;=\; \int_{\{0\}} f(t)\,dt \;=\; \int_{\{0\}} \frac{1}{3}\delta(t)\,dt \;=\; \frac{1}{3}.$$

We now show that if X is a mixed random variable as in (4.3), then $\tilde{f}(x)$ and $\tilde{p}_i$ can be recovered from the cdf of X. From (4.3),

$$F_X(x) \;=\; \int_{-\infty}^{x} \tilde{f}(t)\,dt + \sum_{i:x_i \le x} \tilde{p}_i,$$

and so

$$F_X'(x) \;=\; \tilde{f}(x), \quad x \neq x_i,$$

while

$$F_X(x_i) - F_X(x_i-) \;=\; \tilde{p}_i.$$

Figure 4.7 shows the cdf of a mixed random variable with $\tilde{f}(t) = (0.7/\pi)/(1+t^2)$, $x_1 = 0, x_2 = 3$, $\tilde{p}_1 = 0.21$, and $\tilde{p}_2 = 0.09$.

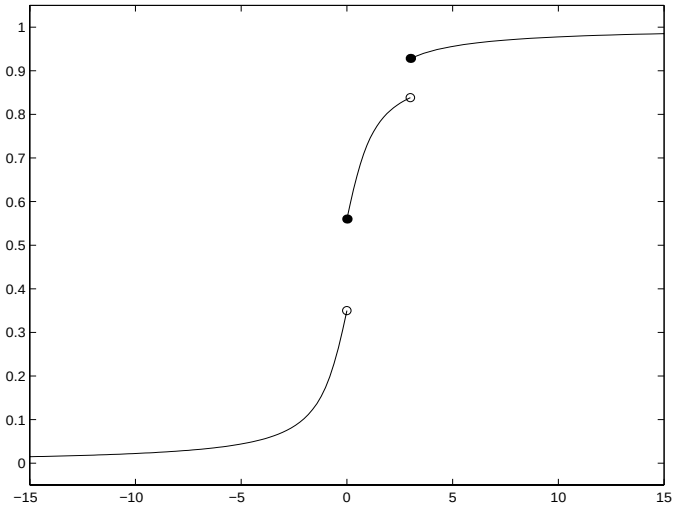

Figure 4.7. Cumulative distribution function of a mixed random variable.

4.5. Functions of Random Variables and Their Cdfs

Most modern systems today are composed of many subsystems in which the output of one system serves as the input to another. When the input to a system or a subsystem is random, so is the output. To evaluate system performance, it is necessary to take into account this randomness. The first step in this process is to find the cdf of the system output if we know the pmf or density of the random input. In many cases, the output will be a mixed random variable with a generalized impulsive density.

We consider systems modeled by real-valued functions $g(x)$. The random system input is a random variable X, and the system output is the random variable $Y = g(X)$. To find $F_Y(y)$, observe that

$$F_Y(y) \;:=\; \wp(Y \le y) \;=\; \wp(g(X) \le y) \;=\; \wp(X \in B_y),$$

where

$$B_y \;:=\; \{x \in \mathbb{R} : g(x) \le y\}.$$

If X has density f_X, then

$$F_Y(y) \;=\; \wp(X \in B_y) \;=\; \int_{B_y} f_X(x)\,dx.$$

The difficulty is to identify the set B_y. However, if we first sketch the function $g(x)$, the problem becomes manageable.

***Example* 4.10.** Suppose g is given by

$$g(x) \;:=\; \begin{cases} 1, & \diamond\!\!\!\!-2 \le x < \diamond\!\!\!\!-1, \\ x^2, & \diamond\!\!\!\!-1 \le x < 0, \\ x, & 0 \le x < 2, \\ 2, & 2 \le x < 3, \\ 0, & \text{otherwise.} \end{cases}$$

Find the inverse image $B = \{x \in \mathbb{R} : g(x) \le y\}$ for $y \ge 2$, $2 > y \ge 1$, $1 > y \ge 0$, and $y < 0$.

Solution. To begin, we sketch g in Figure 4.8. Fix any $y \ge 2$. Since g is upper bounded by 2, $g(x) \le 2 \le y$ for all $x \in \mathbb{R}$. Thus, for $y \ge 2$, $B = \mathbb{R}$.

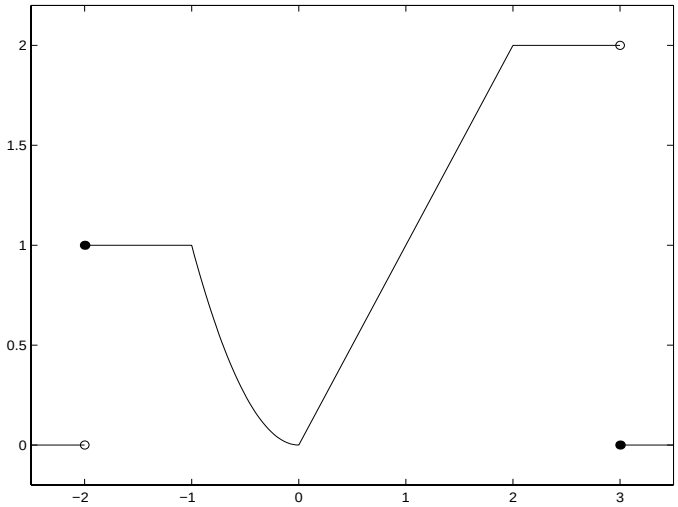

Figure 4.8. The function g from Example 4.10.

Next, fix any y with $2 > y \ge 1$. On the graph of g, draw a horizontal line at level y. In Figure 4.9, the line is drawn at level $y = 1.5$. Next, at the intersection of the horizontal line and the curve g, drop a vertical line to the x-axis. In the figure, this vertical line hits the x-axis at the point marked $\times$.

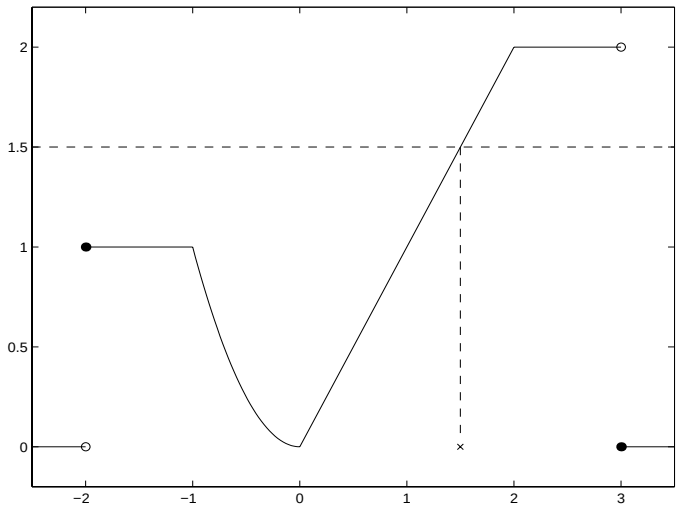

Figure 4.9. Determining B for $1 \le y < 2$.

Clearly, for all x to the left of this point, and for all $x \ge 3$, $g(x) \le y$. To find the x-coordinate of $\times$, we solve $g(x) = y$ for x. For the y-value in question, the formula for $g(x)$ is $g(x) = x$. Hence, the x-coordinate of $\times$ is simply y. Thus, $g(x) \le y \Leftrightarrow x \le y$ or $x \ge 3$, and so,

$$B \;=\; (\Leftrightarrow\infty, y] \cup [3, \infty), \quad 1 \le y < 2.$$

Now fix any y with $1 > y \ge 0$, and draw a horizontal line at level y as before. In Figure 4.10 we have used $y = 1/2$. This time the horizontal line intersects the curve g in two places, and there are two points marked $\times$ on the x-axis. Call the x-coordinate of the left one x_1 and that of the right one x_2. We must solve $g(x_1) = y$, where x_1 is negative and $g(x_1) = x_1^2$. We must also solve $g(x_2) = y$, where $g(x_2) = x_2$. We conclude that $g(x) \le y \Leftrightarrow x < \Leftrightarrow\sqrt{y}$ or $\Leftrightarrow\sqrt{y} \le x \le y$ or $x \ge 3$. Thus,

$$B \;=\; (\Leftrightarrow\infty, \Leftrightarrow\sqrt{y}) \cup [\Leftrightarrow\sqrt{y}, y] \cup [3, \infty), \quad 0 \le y < 1.$$

Note that when $y = 0$, the interval $[0, 0]$ is just the singleton set $\{0\}$.

Finally, since $g(x) \ge 0$ for all $x \in \mathbb{R}$, for $y < 0$, $B = \emptyset$.

***Example* 4.11.** Let g be as in Example 4.10, and suppose that $X \sim$ uniform$[\Leftrightarrow 4, 4]$. Find and sketch the cumulative distribution of $Y = g(X)$. Also find and sketch the density.

Solution. Proceeding as in the two examples above, write $F_Y(y) = \mathcal{P}(Y \le y) = \mathcal{P}(X \in B)$ for the appropriate inverse image $B = \{x \in \mathbb{R} : g(x) \le y\}$.

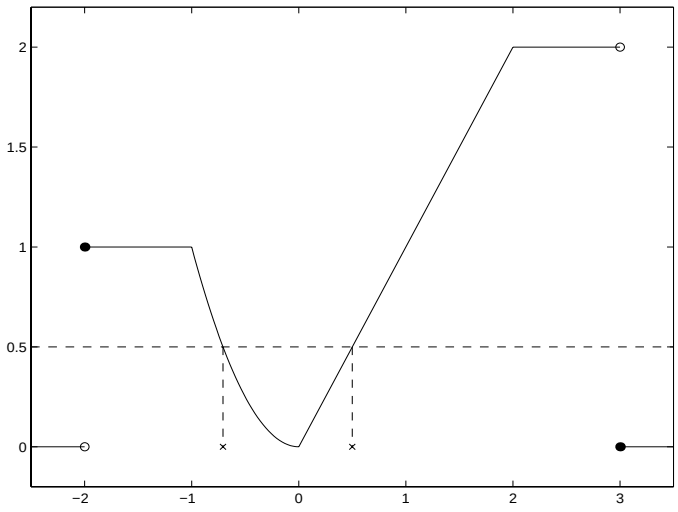

Figure 4.10. Determining B for $0 \le y < 1$.

From Example 4.10,

$$B = \begin{cases} \emptyset, & y < 0, \\ (\Leftarrow\infty, \Leftarrow 2) \cup [\Leftarrow\sqrt{y}, y] \cup [3, \infty), & 0 \le y < 1, \\ (\Leftarrow\infty, y] \cup [3, \infty), & 1 \le y < 2, \\ \mathbb{R}, & y \ge 2. \end{cases}$$

For $y < 0$, $F_Y(y) = 0$. For $0 \le y < 1$,

$$\begin{aligned} F_Y(y) &= \wp(X < \Leftarrow 2) + \wp(\Leftarrow\sqrt{y} \le X \le y) + \wp(X \ge 3) \\ &= \frac{\Leftarrow 2 \Leftarrow (\Leftarrow 4)}{8} + \frac{y \Leftarrow (\Leftarrow\sqrt{y})}{8} + \frac{4 \Leftarrow 3}{8} \\ &= \frac{y + \sqrt{y} + 3}{8}. \end{aligned}$$

For $1 \le y < 2$,

$$\begin{aligned} F_Y(y) &= \wp(X \le y) + \wp(X \ge 3) \\ &= \frac{y \Leftarrow (\Leftarrow 4)}{8} + \frac{4 \Leftarrow 3}{8} \\ &= \frac{y + 5}{8}. \end{aligned}$$

For $y > 2$, $F_Y(y) = \wp(X \in \mathbb{R}) = 1$. Putting all this together,

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ (y + \sqrt{y} + 3)/8, & 0 \le y < 1, \\ (y + 5)/8, & 1 \le y < 2, \\ 1, & y \ge 2. \end{cases}$$

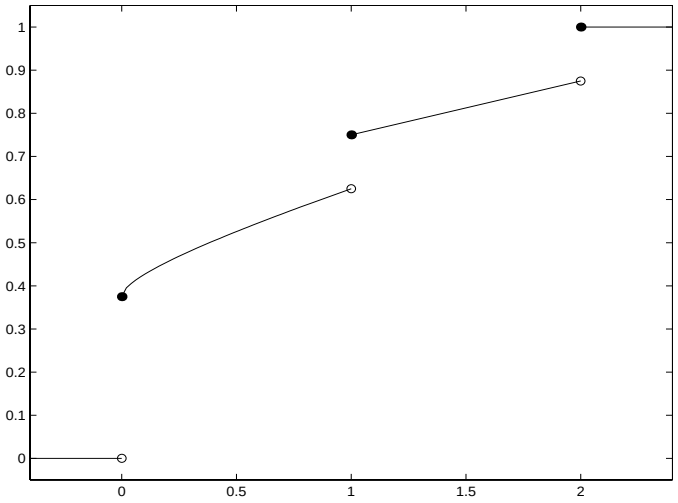

Figure 4.11. Cumulative distribution function $F_Y(y)$ from Example 4.11.

In sketching $F_Y(y)$, we note from the formula that it is 0 for $y < 0$ and 1 for $y \geq 2$. Also from the formula, note that there is a jump discontinuity of $3/8$ at $y = 0$ and a jump of $1/8$ at $y = 1$ and at $y = 2$. See Figure 4.11.

From the observations used in graphing F_Y, we can easily obtain the generalized density,

$$h_Y(y) \;=\; \frac{3}{8}\delta(y) + \frac{1}{8}\delta(y \Leftrightarrow 1) + \frac{1}{8}\delta(y \Leftrightarrow 2) + \tilde{f}_Y(y),$$

where

$$\tilde{f}_Y(y) \;=\; \begin{cases} [1 + 1/(2\sqrt{y})]/8, & 0 < y < 1, \\ 1/8, & 1 < y < 2, \\ 0, & \text{otherwise}, \end{cases}$$

is obtained by differentiating $F_Y(y)$ at non-jump points y. A sketch of h_Y is shown in Figure 4.12.

4.6. Properties of Cdfs

Given an arbitrary real-valued random variable X, its cumulative distribution function is defined by

$$F(x) \;:=\; \wp(X \leq x), \quad \Leftrightarrow\infty < x < \infty.$$

We show below that F satisfies the following eight properties:

(*i*) $0 \leq F(x) \leq 1$.
(*ii*) For $a < b$, $\wp(a < X \leq b) = F(b) \Leftrightarrow F(a)$.

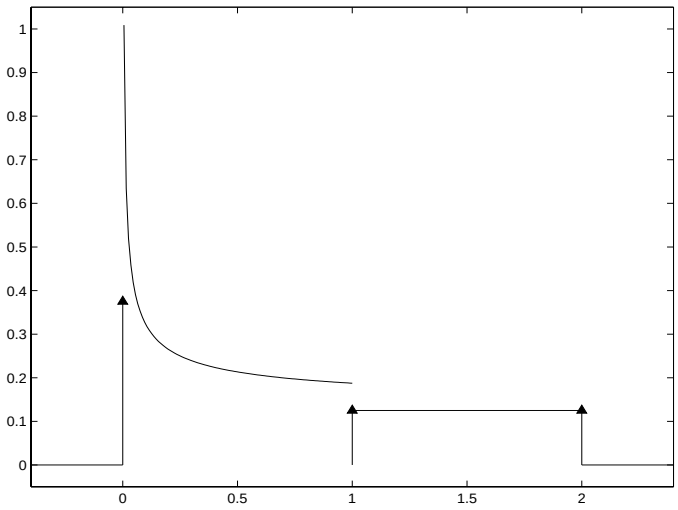

Figure 4.12. Density $h_Y(y)$ from Example 4.11.

(*iii*) F is nondecreasing, i.e., $a \leq b$ implies $F(a) \leq F(b)$.

(*iv*) $\lim_{x \uparrow \infty} F(x) = 1$.

(*v*) $\lim_{x \downarrow -\infty} F(x) = 0$.

(*vi*) $F(x_0+) := \lim_{x \downarrow x_0} F(x) = \wp(X \leq x_0) = F(x_0)$. In other words, F is right-continuous.

(*vii*) $F(x_0\Leftrightarrow) := \lim_{x \uparrow x_0} F(x) = \wp(X < x_0)$.

(*viii*) $\wp(X = x_0) = F(x_0) \Leftrightarrow F(x_0\Leftrightarrow)$.

We also point out that

$$\wp(X > x_0) \;=\; 1 \Leftrightarrow \wp(X \leq x_0) \;=\; 1 \Leftrightarrow F(x_0),$$

and

$$\wp(X \geq x_0) \;=\; 1 \Leftrightarrow \wp(X < x_0) \;=\; 1 \Leftrightarrow F(x_0\Leftrightarrow).$$

If $F(x)$ is continuous at $x = x_0$, i.e., $F(x_0\Leftrightarrow) = F(x_0)$, then this last equation becomes

$$\wp(X \geq x_0) \;=\; 1 \Leftrightarrow F(x_0).$$

Another consequence of the continuity of $F(x)$ at $x = x_0$ is that $\wp(X = x_0) = 0$. Note that if a random variable has a nonimpulsive density, then its cumulative distribution is continuous everywhere.

We now derive the eight properties of cumulative distribution functions.

(*i*) The properties of $\wp$ imply that $F(x) = \wp(X \leq x)$ satisfies $0 \leq F(x) \leq 1$.

(*ii*) First consider the disjoint union $(\Leftrightarrow\infty, b] = (\Leftrightarrow\infty, a] \cup (a, b]$. It then follows that

$$\{X \le b\} \;=\; \{X \le a\} \cup \{a < X \le b\}$$

is a disjoint union of events in Ω. Now write

$$\begin{aligned} F(b) &= \wp(X \le b) \\ &= \wp(\{X \le a\} \cup \{a < X \le b\}) \\ &= \wp(X \le a) + \wp(a < X \le b) \\ &= F(a) + \wp(a < X \le b). \end{aligned}$$

Now subtract $F(a)$ from both sides.

(*iii*) This follows from (*ii*) since $\wp(a < X \le b) \ge 0$.

(*iv*) We prove the simpler result $\lim_{N\to\infty} F(N) = 1$. Starting with

$$\mathbb{R} \;=\; (\Leftrightarrow\infty, \infty) \;=\; \bigcup_{n=1}^{\infty} (\Leftrightarrow\infty, n],$$

we can write

$$\begin{aligned} 1 &= \wp(X \in \mathbb{R}) \\ &= \wp\left(\bigcup_{n=1}^{\infty} \{X \le n\} \right) \\ &= \lim_{N\to\infty} \wp(X \le N), \quad \text{by limit property (1.4)}, \\ &= \lim_{N\to\infty} F(N). \end{aligned}$$

(*v*) We prove the simpler result, $\lim_{N\to\infty} F(\Leftrightarrow N) = 0$. Starting with

$$\emptyset \;=\; \bigcap_{n=1}^{\infty} (\Leftrightarrow\infty, \Leftrightarrow n],$$

we can write

$$\begin{aligned} 0 &= \wp(X \in \emptyset) \\ &= \wp\left(\bigcap_{n=1}^{\infty} \{X \le \Leftrightarrow n\} \right) \\ &= \lim_{N\to\infty} \wp(X \le \Leftrightarrow N), \quad \text{by limit property (1.5)}, \\ &= \lim_{N\to\infty} F(\Leftrightarrow N). \end{aligned}$$

(*vi*) We prove the simpler result, $\wp(X \le x_0) = \lim_{N\to\infty} F(x_0 + \frac{1}{N})$. Starting with

$$(\Leftrightarrow\infty, x_0] \;=\; \bigcap_{n=1}^{\infty} (\Leftrightarrow\infty, x_0 + \tfrac{1}{n}],$$

we can write

$$\begin{aligned} \wp(X \le x_0) &= \wp\Big(\bigcap_{n=1}^{\infty} \{X \le x_0 + \tfrac{1}{n}\} \Big) \\ &= \lim_{N\to\infty} \wp(X \le x_0 + \tfrac{1}{N}), \quad \text{by (1.5)}, \\ &= \lim_{N\to\infty} F(x_0 + \tfrac{1}{N}). \end{aligned}$$

(*vii*) We prove the simpler result, $\wp(X < x_0) = \lim_{N\to\infty} F(x_0 - \frac{1}{N})$. Starting with

$$(-\infty, x_0) = \bigcup_{n=1}^{\infty} (-\infty, x_0 - \tfrac{1}{n}],$$

we can write

$$\begin{aligned} \wp(X < x_0) &= \wp\Big(\bigcup_{n=1}^{\infty} \{X \le x_0 - \tfrac{1}{n}\} \Big) \\ &= \lim_{N\to\infty} \wp(X \le x_0 - \tfrac{1}{N}), \quad \text{by (1.4)}, \\ &= \lim_{N\to\infty} F(x_0 - \tfrac{1}{N}). \end{aligned}$$

(*viii*) First consider the disjoint union $(-\infty, x_0] = (-\infty, x_0) \cup \{x_0\}$. It then follows that

$$\{X \le x_0\} = \{X < x_0\} \cup \{X = x_0\}$$

is a disjoint union of events in Ω. Using Property (*vii*), it follows that

$$F(x_0) = F(x_0-) + \wp(X = x_0).$$

4.7. The Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. with common mean m and common variance σ^2. There are many cases for which we know the probability mass function or density of

$$\sum_{i=1}^{n} X_i.$$

For example, if the X_i are Bernoulli, binomial, Poisson, gamma, or Gaussian, we know the cdf of the sum (see Example 2.20 or Problem 22 in Chapter 2 and Problem 46 in Chapter 3). Note that the exponential and chi-squared are special cases of the gamma (see Problem 12 in Chapter 3). In general, however, finding the cdf of a sum of i.i.d. random variables is not possible. Fortunately, we have the following approximation.

Central Limit Theorem (CLT). *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with finite mean m and finite variance σ^2. If*

$$Y_n := \frac{M_n - m}{\sigma/\sqrt{n}} \quad \text{and} \quad M_n := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

then

$$\lim_{n\to\infty} F_{Y_n}(y) \;=\; \Phi(y).$$

To get some idea of how large n should be, would like to compare $F_{Y_n}(y)$ and $\Phi(y)$ in cases where F_{Y_n} is known. To do this, we need the following result.

***Example* 4.12.** Show that if G_n is the cdf of $\sum_{i=1}^n X_i$, then

$$F_{Y_n}(y) \;=\; G_n(y\sigma\sqrt{n}+nm). \tag{4.6}$$

Solution. Write

$$\begin{aligned} F_{Y_n}(y) &= \wp\left(\frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}} \le y\right) \\ &= \wp(M_n \Leftrightarrow m \le y\sigma/\sqrt{n}\,) \\ &= \wp(M_n \le y\sigma/\sqrt{n}+m) \\ &= sP\left(\sum_{i=1}^n X_i \le y\sigma\sqrt{n}+nm\right) \\ &= G_n(y\sigma\sqrt{n}+nm). \end{aligned}$$

When the X_i are exp(1), G_n is the Erlang$(30,1)$ cdf given in Problem 12(c) in Chapter 3. In Figure 4.13 we plot $F_{Y_{30}}(y)$ (dashed line) and the $N(0,1)$ cdf $\Phi(y)$ (solid line).

A typical calculation using the central limit theorem is as follows. To approximate

$$\wp\left(\sum_{i=1}^n X_i > t\right),$$

write

$$\begin{aligned} \wp\left(\sum_{i=1}^n X_i > t\right) &= \wp\left(\frac{1}{n}\sum_{i=1}^n X_i > \frac{t}{n}\right) \\ &= \wp(M_n > t/n) \\ &= \wp(M_n \Leftrightarrow m > t/n \Leftrightarrow m) \\ &= \wp\left(\frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}} > \frac{t/n \Leftrightarrow m}{\sigma/\sqrt{n}}\right) \\ &= \wp\left(Y_n > \frac{t/n \Leftrightarrow m}{\sigma/\sqrt{n}}\right) \\ &= 1 \Leftrightarrow F_{Y_n}\left(\frac{t/n \Leftrightarrow m}{\sigma/\sqrt{n}}\right) \\ &\approx 1 \Leftrightarrow \Phi\left(\frac{t/n \Leftrightarrow m}{\sigma/\sqrt{n}}\right). \end{aligned}$$

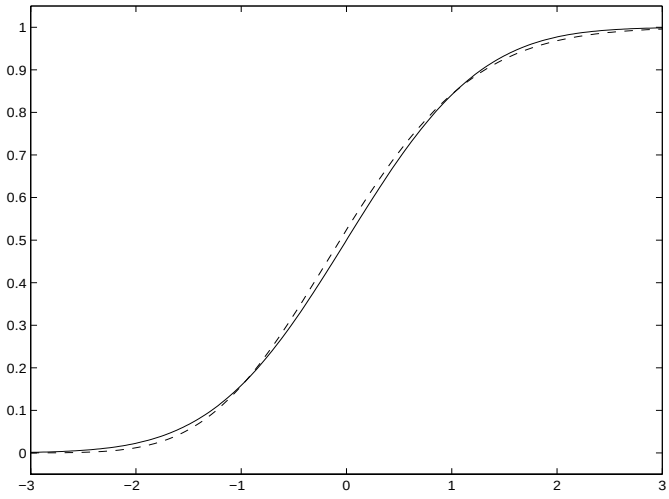

Figure 4.13. Illustration of the central limit theorem when the X_i are exponential with parameter 1. The dashed line is $F_{Y_{30}}$, and the solid line is the standard normal cumulative distribution, Φ.

***Example* 4.13.** A certain digital communication link has bit-error probability p. Use the central limit theorem to approximate the probability that in transmitting a word of n bits, more than k bits are received incorrectly.

Solution. Let $X_i = 1$ if bit i is received in error, and $X_i = 0$ otherwise. We assume the X_i are independent Bernoulli(p) random variables. The number of errors in n bits is $\sum_{i=1}^{n} X_i$. We must compute

$$\wp\left(\sum_{i=1}^{n} X_i > k\right).$$

Using the approximation above in which $t = k$, $m = p$, and $\sigma^2 = p(1 \Leftrightarrow p)$, we have

$$\wp\left(\sum_{i=1}^{n} X_i > k\right) \approx 1 \Leftrightarrow \Phi\left(\frac{k/n \Leftrightarrow p}{\sqrt{p(1 \Leftrightarrow p)/n}}\right).$$

Derivation of the Central Limit Theorem

It is instructive to consider first the following special case, which illustrates the key steps of the general derivation. Suppose that the X_i are i.i.d. Laplace with parameter $\lambda = \sqrt{2}$. Then $m = 0$, $\sigma^2 = 1$, and

$$Y_n = \frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}} = \sqrt{n} M_n = \frac{\sqrt{n}}{n} \sum_{i=1}^{n} X_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

The characteristic function of Y_n is

$$\varphi_{Y_n}(\nu) \;=\; \mathsf{E}[e^{j\nu Y_n}] \;=\; \mathsf{E}\left[e^{j(\nu/\sqrt{n})\sum_{i=1}^{n} X_i}\right] \;=\; \prod_{i=1}^{n} \mathsf{E}\left[e^{j(\nu/\sqrt{n})X_i}\right].$$

Of course, $\mathsf{E}\left[e^{j(\nu/\sqrt{n})X_i}\right] = \varphi_{X_i}\left(\nu/\sqrt{n}\,\right)$, where, for the Laplace$\left(\sqrt{2}\,\right)$ random variable X_i,

$$\varphi_{X_i}(\nu) \;=\; \frac{2}{2+\nu^2} \;=\; \frac{1}{1+\nu^2/2}.$$

Thus,

$$\mathsf{E}[e^{j(\nu/\sqrt{n})X_i}] \;=\; \varphi_{X_i}\left(\frac{\nu}{\sqrt{n}}\right) \;=\; \frac{1}{1+\frac{\nu^2/2}{n}},$$

and

$$\varphi_{Y_n}(\nu) \;=\; \left(\frac{1}{1+\frac{\nu^2/2}{n}}\right)^n \;=\; \frac{1}{\left(1+\frac{\nu^2/2}{n}\right)^n}.$$

We now use the fact that for any number ξ,

$$\left(1+\frac{\xi}{n}\right)^n \;\to\; e^{\xi}.$$

It follows that

$$\varphi_{Y_n}(\nu) \;=\; \frac{1}{\left(1+\frac{\nu^2/2}{n}\right)^n} \;\to\; \frac{1}{e^{\nu^2/2}} \;=\; e^{-\nu^2/2},$$

which is the characteristic function of an $N(0,1)$ random variable.

We now turn to the derivation in the general case. Write

$$\begin{aligned} Y_n &= \frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}} \\ &= \frac{\sqrt{n}}{\sigma}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \Leftrightarrow m\right] \\ &= \frac{\sqrt{n}}{\sigma}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i \Leftrightarrow m)\right] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i \Leftrightarrow m}{\sigma}\right). \end{aligned}$$

Now let $Z_i := (X_i \Leftrightarrow m)/\sigma$. Then

$$Y_n \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i,$$

where the Z_i are zero mean and unit variance. Since the X_i are i.i.d., so are the Z_i. Let $\varphi_Z(\nu) := \mathsf{E}[e^{j\nu Z_i}]$ denote their common characteristic function. We can write the characteristic function of Y_n as

$$\begin{aligned}
\varphi_{Y_n}(\nu) &:= \mathsf{E}[e^{j\nu Y_n}] \\
&= \mathsf{E}\left[\exp\left(j\frac{\nu}{\sqrt{n}}\sum_{i=1}^{n} Z_i\right)\right] \\
&= \mathsf{E}\left[\prod_{i=1}^{n}\exp\left(j\frac{\nu}{\sqrt{n}}Z_i\right)\right] \\
&= \prod_{i=1}^{n}\mathsf{E}\left[\exp\left(j\frac{\nu}{\sqrt{n}}Z_i\right)\right] \\
&= \prod_{i=1}^{n}\varphi_Z\left(\frac{\nu}{\sqrt{n}}\right) \\
&= \varphi_Z\left(\frac{\nu}{\sqrt{n}}\right)^n.
\end{aligned}$$

Now recall that for any complex ξ,

$$e^{\xi} = 1 + \xi + \frac{1}{2}\xi^2 + R(\xi).$$

Thus,

$$\begin{aligned}
\varphi_Z\left(\frac{\nu}{\sqrt{n}}\right) &= \mathsf{E}[e^{j(\nu/\sqrt{n})Z_i}] \\
&= \mathsf{E}\left[1 + j\frac{\nu}{\sqrt{n}}Z_i + \frac{1}{2}\left(j\frac{\nu}{\sqrt{n}}Z_i\right)^2 + R\left(j\frac{\nu}{\sqrt{n}}Z_i\right)\right].
\end{aligned}$$

Since Z_i is zero mean and unit variance,

$$\varphi_Z\left(\frac{\nu}{\sqrt{n}}\right) = 1 \Leftrightarrow \frac{1}{2}\cdot\frac{\nu^2}{n} + \mathsf{E}\left[R\left(j\frac{\nu}{\sqrt{n}}Z_i\right)\right].$$

It can be shown that the last term on the right is asymptotically negligible [4, pp. 357–358], and so

$$\varphi_Z\left(\frac{\nu}{\sqrt{n}}\right) \approx 1 \Leftrightarrow \frac{\nu^2/2}{n}.$$

We now have

$$\varphi_{Y_n}(\nu) = \varphi_Z\left(\frac{\nu}{\sqrt{n}}\right)^n \approx \left(1 \Leftrightarrow \frac{\nu^2/2}{n}\right)^n \to e^{-\nu^2/2},$$

which is the $N(0,1)$ characteristic function. Since the characteristic function of Y_n converges to the $N(0,1)$ characteristic function, it follows that $F_{Y_n}(y) \to \Phi(y)$ [4, p. 349, Theorem 26.3].

***Example* 4.14.** In the derivation of the central limit theorem, we found that $\varphi_{Y_n}(\nu) \to e^{-\nu^2/2}$ as $n \to \infty$. We can use this result to give a simple derivation of **Stirling's formula**,

$$n! \;\approx\; \sqrt{2\pi}\, n^{n+1/2} e^{-n}.$$

Solution. Since $n! = n(n \Leftrightarrow 1)!$, it suffices to show that

$$(n \Leftrightarrow 1)! \;\approx\; \sqrt{2\pi}\, n^{n-1/2} e^{-n}.$$

To this end, let $X_1, \ldots, X_n$ be i.i.d. exp(1). Then by Problem 46(c) in Chapter 3, $\sum_{i=1}^n X_i$ has the n-Erlang density,

$$g_n(x) \;=\; \frac{x^{n-1} e^{-x}}{(n \Leftrightarrow 1)!}, \quad x \geq 0.$$

Also, the mean and variance of X_i are both one. Differentiating (4.6) (with $m = \sigma = 1$) yields

$$f_{Y_n}(y) \;=\; g_n(y\sqrt{n} + n)\sqrt{n}.$$

Note that $f_{Y_n}(0) = n^{n-1/2} e^{-n}/(n \Leftrightarrow 1)!$. On the other hand, by inverting the characteristic function of Y_n, we can also write

$$\begin{aligned} f_{Y_n}(0) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_{Y_n}(\nu)\, d\nu \\ &\approx \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\nu^2/2}\, d\nu, \quad \text{by the CLT,} \\ &= 1/\sqrt{2\pi}. \end{aligned}$$

It follows that $(n \Leftrightarrow 1)! \approx \sqrt{2\pi}\, n^{n-1/2} e^{-n}$.

Remark. A more precise version of Stirling's formula is [14, pp. 50–53]

$$\sqrt{2\pi}\, n^{n+1/2} e^{-n+1/(12n+1)} \;< n! \;< \sqrt{2\pi}\, n^{n+1/2} e^{-n+1/(12n)}.$$

4.8. Problems

Problems §4.1: Continuous Random Variables

1. Find the cumulative distribution function $F_X(x)$ of an exponential random variable X with parameter λ. Sketch the graph of F_X when $\lambda = 1$.

2. The **Rayleigh** density with parameter λ is defined by

$$f(x) \;:=\; \begin{cases} \dfrac{x}{\lambda^2} e^{-(x/\lambda)^2/2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

   Find the cumulative distribution function.

*3. The **Maxwell** density with parameter λ is defined by

$$f(x) := \begin{cases} \sqrt{\frac{2}{\pi}} \frac{x^2}{\lambda^3} e^{-(x/\lambda)^2/2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Show that the cdf $F(x)$ can be expressed in terms of the standard normal cdf $\Phi(x)$ defined in (4.1).

4. If Z has density $f_Z(z)$ and $Y = e^Z$, find $f_Y(y)$.

5. If $X \sim \text{uniform}(0,1)$, find the density of $Y = \ln(1/X)$.

6. Let $X \sim \text{Weibull}(p, \lambda)$. Find the density of $Y = \lambda X^p$.

7. The input to a squaring circuit is a Gaussian random variable X with mean zero and variance one. Show that the output $Y = X^2$ has the chi-squared density with one degree of freedom,

$$f_Y(y) = \frac{e^{-y/2}}{\sqrt{2\pi y}}, \quad y > 0.$$

8. If the input to the squaring circuit of Problem 7 includes a fixed bias, say m, then the output is given by $Y = (X+m)^2$, where again $X \sim N(0,1)$. Show that Y has the noncentral chi-squared density with one degree of freedom and noncentrality parameter m^2,

$$f_Y(y) = \frac{e^{-(y+m^2)/2}}{\sqrt{2\pi y}} \cdot \frac{e^{m\sqrt{y}} + e^{-m\sqrt{y}}}{2}, \quad y > 0.$$

Note that if $m = 0$, we recover the result of Problem 7.

9. Let $X_1, \ldots, X_n$ be independent with common cumulative distribution function $F(x)$. Let us define $X_{\max} := \max(X_1, \ldots, X_n)$ and $X_{\min} := \min(X_1, \ldots, X_n)$. Express the cumulative distributions of $X_{\max}$ and $X_{\min}$ in terms of $F(x)$. *Hint:* Example 2.7 may be helpful.

10. If X and Y are independent $\exp(\lambda)$ random variables, find $\mathsf{E}[\max(X,Y)]$.

11. *Digital Communication System.* The received voltage in a digital communication system is $Z = X + Y$, where $X \sim \text{Bernoulli}(p)$ is a random message, and $Y \sim N(0,1)$ is a Gaussian noise voltage. Assuming X and Y are independent, find the conditional cdf $F_{Z|X}(z|i)$ for $i = 0, 1$, the cdf $F_Z(z)$, and the density $f_Z(z)$.

12. *Fading Channel.* Let X and Y be as in the preceding problem, but now suppose $Z = X/A + Y$, where A, X, and Y are independent, and A takes the values 1 and 2 with equal probability. Find the conditional cdf $F_{Z|A,X}(z|a,i)$ for $a = 1, 2$ and $i = 0, 1$.

13. *Generalized Rayleigh Densities.* Let Y_n be chi-squared with n degrees of freedom as defined in Problem 12 of Chapter 3. Put $Z_n := \sqrt{Y_n}$.

    (a) Express the cdf of Z_n in terms of the cdf of Y_n.

    (b) Find the density of Z_1.

    (c) Show that Z_2 has a Rayleigh density, as defined in Problem 2, with $\lambda = 1$.

    (d) Show that Z_3 has a Maxwell density, as defined in Problem 3, with $\lambda = 1$.

    (e) Show that Z_{2m} has a **Nakagami-*m*** density

$$f(z) := \begin{cases} \dfrac{2}{2^m, (m)} \dfrac{z^{2m-1}}{\lambda^{2m}} e^{-(z/\lambda)^2/2}, & z \ge 0, \\ 0, & z < 0, \end{cases}$$

    with $\lambda = 1$.

Remark. For the general chi-squared random variable Y_n, it is not necessary that n be an integer. However, if n is a positive integer, and if $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$, then the X_i^2 are chi-squared with one degree of freedom by Problem 7, and $Y_n := X_1^2 + \cdots + X_n^2$ is chi-squared with n degrees of freedom by Problem 46(c) in Chapter 3. Hence, the above densities usually arise from taking the square root of the sum of squares of standard normal random variables. For example, (X_1, X_2) can be regarded as a random point in the plane whose horizontal and vertical coordinates are independent $N(0,1)$. The distance of this point from the origin is $\sqrt{X_1^2 + X_2^2} = Z_2$, which is a Rayleigh random variable. As another example, consider an ideal gas. The velocity of a given particle is obtained by adding up the results of many collisions with other particles. By the central limit theorem (Section 4.7), each component of the given particle's velocity vector, say (X_1, X_2, X_3) should be i.i.d. $N(0,1)$. The speed of the particle is $\sqrt{X_1^2 + X_2^2 + X_3^2} = Z_3$, which has the Maxwell density. When the Nakagami-m density is used as a model for fading in wireless communication channels, m is often not an integer.

14. *Generalized Gamma Densities.*

    (a) Let $X \sim \text{gamma}(p, 1)$, and put $Y := X^{1/q}$. Show that

$$f_Y(y) = \frac{q y^{pq-1} e^{-y^q}}{, (p)}, \quad y > 0.$$

    (b) If in part (a) we replace p with p/q, we find that

$$f_Y(y) = \frac{q y^{p-1} e^{-y^q}}{, (p/q)}, \quad y > 0.$$

If we introduce a scale parameter $\lambda > 0$, we have the **generalized gamma** density [46]. More precisely, we say that $Y \sim$ g-gamma(p, λ, q) if Y has density

$$f_Y(y) \;=\; \frac{\lambda q(\lambda y)^{p-1} e^{-(\lambda y)^q}}{, (p/q)}, \quad y > 0.$$

Clearly, g-gamma$(p, \lambda, 1) =$ gamma(p, λ), which includes the Erlang and the chi-squared as special cases. Show that

(i) g-gamma$(p, \lambda^{1/p}, p) =$ Weibull(p, λ).

(ii) g-gamma$(2, 1/(\sqrt{2}\lambda), 2)$ is the Rayleigh density defined in Problem 2.

(iii) g-gamma$(3, 1/(\sqrt{2}\lambda), 2)$ is the Maxwell density defined in Problem 3.

(c) If $Y \sim$ g-gamma(p, λ, q), show that

$$\mathsf{E}[Y^n] \;=\; \frac{, \big((n+p)/q\big)}{, (p/q)\lambda^n},$$

and conclude that

$$M_Y(s) \;=\; \sum_{n=0}^{\infty} \frac{s^n}{n!} \cdot \frac{, \big((n+p)/q\big)}{, (p/q)\lambda^n}.$$

⋆15. In the analysis of communication systems, one is often interested in $\wp(X > x) = 1 \Leftrightarrow F_X(x)$ for some voltage threshold x. We call $F_X^c(x) := 1 \Leftrightarrow F_X(x)$ the **complementary cumulative distribution function** (ccdf) of X. Of particular interest is the ccdf of the standard normal, which is often denoted by

$$Q(x) \;:=\; 1 \Leftrightarrow \Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt.$$

Using the hints below, show that for $x > 0$,

$$\frac{e^{-x^2/2}}{\sqrt{2\pi}} \left(\frac{1}{x} \Leftrightarrow \frac{1}{x^3} \right) \;<\; Q(x) \;<\; \frac{e^{-x^2/2}}{x\sqrt{2\pi}}.$$

Hints: To derive the upper bound, apply integration by parts to

$$\int_x^\infty \frac{1}{t} \cdot t e^{-t^2/2}\, dt,$$

and then drop the new integral term (which is positive),

$$\int_x^\infty \frac{1}{t^2} e^{-t^2/2}\, dt.$$

If you do *not* drop the above term, you can derive the lower bound by applying integration by parts one more time (after dividing and multiplying by t again) and then dropping the final integral.

*16. Let $C_k(x)$ denote the chi-squared cdf with k degrees of freedom. Show that the noncentral chi-squared cdf with k degrees of freedom and noncentrality parameter λ^2 is given by (recall Problem 54 in Chapter 3)

$$C_{k,\lambda^2}(x) = \sum_{n=0}^{\infty} \frac{(\lambda^2/2)^n e^{-\lambda^2/2}}{n!} C_{2n+k}(x).$$

Remark. Note that in MATLAB, $C_k(x) =$ `chi2cdf(x,k)`, and $C_{k,\lambda^2}(x) =$ `ncx2cdf(x,k,lambda^2)`.

*17. *Generalized Rice or Noncentral Rayleigh Densities.* Let Y_n be noncentral chi-squared with n degrees of freedom and noncentrality parameter m^2 as defined in Problem 54 in Chapter 3. (In general, n need not be an integer, but if it is, and if $X_1, \ldots, X_n$ are independent normal random variables with $X_i \sim N(m_i, 1)$, then by Problem 8, X_i^2 is noncentral chi-squared with one degree of freedom and noncentrality parameter m_i^2, and by Problem 54 in Chapter 3, $X_1^2 + \cdots + X_n^2$ is noncentral chi-squared with n degrees of freedom and noncentrality parameter $m^2 = m_1^2 + \cdots + m_n^2$.)

(a) Show that $Z_n := \sqrt{Y_n}$ has the **generalized Rice** density,

$$f_{Z_n}(z) = \frac{z^{n/2}}{m^{n/2-1}} e^{-(m^2+z^2)/2} I_{n/2-1}(mz), \quad z > 0,$$

where I_ν is the modified **Bessel function** of the first kind, order ν,

$$I_\nu(x) := \sum_{\ell=0}^{\infty} \frac{(x/2)^{2\ell+\nu}}{\ell!\,,\,(\ell+\nu+1)}.$$

Remark. In MATLAB, $I_\nu(x) =$ `besseli(nu,x)`.

(b) Show that Z_2 has the original Rice density,

$$f_{Z_2}(z) = z e^{-(m^2+z^2)/2} I_0(mz), \quad z > 0.$$

(c) Show that

$$f_{Y_n}(y) = \frac{1}{2}\Big(\frac{\sqrt{y}}{m}\Big)^{n/2-1} e^{-(m^2+y)/2} I_{n/2-1}(m\sqrt{y}), \quad y > 0,$$

giving a closed-form expression for the noncentral chi-squared density. Recall that you already have a closed-form expression for the moment generating function and characteristic function of a noncentral chi-squared random variable (see Problem 54(b) in Chapter 3).

Remark. Note that in MATLAB, the cdf of Y_n is given by $F_{Y_n}(y) =$ `ncx2cdf(y,n,m^2)`.

(d) Denote the complementary cumulative distribution of Z_n by

$$F^c_{Z_n}(z) \;:=\; \wp(Z_n > z) \;=\; \int_z^\infty f_{Z_n}(t)\,dt.$$

Show that

$$F^c_{Z_n}(z) \;=\; \Big(\frac{z}{m}\Big)^{n/2-1} e^{-(m^2+z^2)/2} I_{n/2-1}(mz) + F^c_{Z_{n-2}}(z).$$

Hint: Use integration by parts; you will need the easily-verified fact that

$$\frac{d}{dx}\Big(x^\nu I_\nu(x)\Big) \;=\; x^\nu I_{\nu-1}(x).$$

(e) The complementary cdf of Z_2, $F^c_{Z_2}(z)$, is known as the **Marcum Q function**,

$$Q(m,z) \;:=\; \int_z^\infty t e^{-(m^2+t^2)/2} I_0(mt)\,dt.$$

Show that if $n \geq 4$ is an even integer, then

$$F^c_{Z_n}(z) = Q(m,z) + e^{-(m^2+z^2)/2} \sum_{k=1}^{n/2-1} \Big(\frac{z}{m}\Big)^k I_k(mz).$$

(f) Show that $Q(m,z) = \widetilde{Q}(m,z)$, where

$$\widetilde{Q}(m,z) \;:=\; e^{-(m^2+z^2)/2} \sum_{k=0}^{\infty} (m/z)^k I_k(mz).$$

Hint: [19, p. 450] Show that $Q(0,z) = \widetilde{Q}(0,z) = e^{-z^2/2}$. It then suffices to prove that

$$\frac{\partial}{\partial m} Q(m,z) \;=\; \frac{\partial}{\partial m} \widetilde{Q}(m,z).$$

To this end, use the derivative formula in the hint of part (d) to show that

$$\frac{\partial}{\partial m} \widetilde{Q}(m,z) \;=\; z e^{-(m^2+z^2)/2} I_1(mz).$$

Now take the same partial derivative of $Q(m,z)$ as defined in part (e), and then use integration by parts on the term involving I_1.

⋆18. *Properties of Modified Bessel Functions.*

(a) Use the power-series definition of $I_\nu(x)$ in the preceding problem to show that

$$I'_\nu(x) \;=\; \tfrac{1}{2}[I_{\nu-1}(x) + I_{\nu+1}(x)]$$

and that

$$I_{\nu-1}(x) \Leftrightarrow I_{\nu+1}(x) \;=\; 2(\nu/x)\, I_\nu(x).$$

Note that the second identity implies the recursion,

$$I_{\nu+1}(x) \;=\; I_{\nu-1}(x) \Leftrightarrow 2(\nu/x)\, I_\nu(x).$$

Hence, once $I_0(x)$ and $I_1(x)$ are known, $I_n(x)$ can be computed for $n = 2, 3, \ldots$.

(b) Parts (b) and (c) of this problem are devoted to showing that for integers $n \geq 0$,

$$I_n(x) \;=\; \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{x\cos\theta} \cos(n\theta)\, d\theta.$$

To this end, denote the above integral by $\tilde{I}_n(x)$. Use integration by parts and a trigonometric identity to show that

$$\tilde{I}_n(x) \;=\; \frac{x}{2n}[\tilde{I}_{n-1}(x) \Leftrightarrow \tilde{I}_{n+1}(x)].$$

Hence, in Part (c) it will be enough to show that $\tilde{I}_0(x) = I_0(x)$ and $\tilde{I}_1(x) = I_1(x)$.

(c) From the integral definition of $\tilde{I}_n(x)$, it is clear that $\tilde{I}_0'(x) = \tilde{I}_1(x)$. It is also easy to see from the power series for $I_0(x)$ that $I_0'(x) = I_1(x)$. Hence, it is enough to show that $\tilde{I}_0(x) = I_0(x)$. Since the integrand defining $\tilde{I}_0(x)$ is even,

$$\tilde{I}_0(x) \;=\; \frac{2}{\pi} \int_0^{\pi} e^{x\cos\theta}\, d\theta.$$

Show that

$$\tilde{I}_0(x) \;=\; \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} e^{-x\sin t}\, dt.$$

Then use the power series $e^\xi = \sum_{k=0}^{\infty} \xi^k/k!$ in the above integrand and integrate term by term. Then use the results of Problems 15 and 11 of Chapter 3 to show that $\tilde{I}_0(x) = I_0(x)$.

(d) Use the integral formula for $I_n(x)$ to show that

$$I_n'(x) \;=\; \tfrac{1}{2}[I_{n-1}(x) + I_{n+1}(x)].$$

Problems §4.2: Reliability

19. The lifetime T of a Model n Internet router has an Erlang$(n, 1)$ density, $f_T(t) = t^{n-1}e^{-t}/(n \Leftrightarrow 1)!$.

(a) What is the router's mean time to failure?

(b) Show that the reliability of the router after t time units of operation is

$$R(t) \;=\; \sum_{k=0}^{n-1} \frac{t^k}{k!} e^{-t}.$$

(c) Find the failure rate (known as the **Erlang** failure rate).

20. A certain device has the **Weibull** failure rate

$$r(t) \;=\; \lambda p t^{p-1}, \quad t > 0.$$

(a) Sketch the failure rate for $\lambda = 1$ and the cases $p = 1/2$, $p = 1$, $p = 3/2$, $p = 2$, and $p = 3$.

(b) Find the reliability $R(t)$.

(c) Find the mean time to failure.

(d) Find the density $f_T(t)$.

21. A certain device has the **Pareto** failure rate

$$r(t) \;=\; \begin{cases} p/t, & t \geq t_0, \\ 0, & t < t_0. \end{cases}$$

(a) Find the reliability $R(t)$ for $t \geq 0$.

(b) Sketch $R(t)$ if $t_0 = 1$ and $p = 2$.

(c) Find the mean time to failure if $p > 1$.

(d) Find the Pareto density $f_T(t)$.

22. A certain device has failure rate $r(t) = t^2 \Leftrightarrow 2t + 2$ for $t \geq 0$.

(a) Sketch $r(t)$ for $t \geq 0$.

(b) Find the corresponding density $f_T(t)$ in closed form (no integrals).

23. Suppose that the lifetime T of a device is uniformly distributed on the interval $[1, 2]$.

(a) Find and sketch the reliability $R(t)$ for $t \geq 0$.

(b) Find the failure rate $r(t)$ for $0 \leq t < 2$.

(c) Find the mean time to failure.

24. Consider a system composed of two devices with respective lifetimes T_1 and T_2. Let T denote the lifetime of the composite system. Suppose that the system operates properly if and only if *both* devices are functioning. In other words, $T > t$ if and only if $T_1 > t$ *and* $T_2 > t$. Express the reliability of the overall system $R(t)$ in terms of $R_1(t)$ and $R_2(t)$, where $R_1(t)$ and $R_2(t)$ are the reliabilities of the individual devices. Assume T_1 and T_2 are independent.

25. Consider a system composed of two devices with respective lifetimes T_1 and T_2. Let T denote the lifetime of the composite system. Suppose that the system operates properly if and only if *at least one of* the devices is functioning. In other words, $T > t$ if and only if $T_1 > t$ *or* $T_2 > t$. Express the reliability of the overall system $R(t)$ in terms of $R_1(t)$ and $R_2(t)$, where $R_1(t)$ and $R_2(t)$ are the reliabilities of the individual devices. Assume T_1 and T_2 are independent.

26. Let Y be a nonnegative random variable. Show that

$$\mathsf{E}[Y^n] \;=\; \int_0^\infty n y^{n-1} \wp(Y > y)\, dy.$$

Hint: Put $T = Y^n$ in (4.2).

Problems §4.3: Discrete Random Variables

27. Let $X \sim \text{binomial}(n, p)$ with $n = 4$ and $p = 3/4$. Sketch the graph of the cumulative distribution function of X, $F_X(x)$.

Problems §4.4: Mixed Random Variables

28. A random variable X has generalized density

$$f(t) \;=\; \tfrac{1}{3}e^{-t}u(t) + \tfrac{1}{2}\delta(t) + \tfrac{1}{6}\delta(t \Leftrightarrow 1),$$

where u is the unit step function defined in Section 2.1, and δ is the Dirac delta function defined in Section 4.4.

(a) Sketch $f(t)$.

(b) Compute $\wp(X = 0)$ and $\wp(X = 1)$.

(c) Compute $\wp(0 < X < 1)$ and $\wp(X > 1)$.

(d) Use your above results to compute $\wp(0 \le X \le 1)$ and $\wp(X \ge 1)$.

(e) Compute $\mathsf{E}[X]$.

29. If X has generalized density $f(t) = \frac{1}{2}[\delta(t) + I_{(0,1]}(t)]$, evaluate $\mathsf{E}[X]$ and $\wp(X = 0 | X \le 1/2)$.

30. Let X have cdf

$$F_X(x) \;=\; \begin{cases} 0, & x < 0, \\ x^2, & 0 \le x < 1/2, \\ x, & 1/2 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

Show that $\mathsf{E}[X] = 7/12$.

31. Sketch the graph of the cumulative distribution function of the random variable in Example 4.9. Also sketch corresponding impulsive density function.

*32. A certain computer monitor contains a loose connection. The connection is loose with probability 1/2. When the connection is loose, the monitor displays a blank screen (brightness=0). When the connection is not loose, the brightness is uniformly distributed on $(0,1]$. Let X denote the observed brightness. Find formulas and plot the cdf and generalized density of X.

Problems §4.5: Functions of Random Variables and Their Cdfs

33. Let $\Theta \sim \text{uniform}[-\pi, \pi]$, and put $X := \cos\Theta$ and $Y := \sin\Theta$.

    (a) Show that $F_X(x) = 1 - \frac{1}{\pi}\cos^{-1} x$ for $-1 \le x \le 1$.

    (b) Show that $F_Y(y) = \frac{1}{2} + \frac{1}{\pi}\sin^{-1} y$ for $-1 \le y \le 1$.

    (c) Use the identity $\sin(\pi/2 - \theta) = \cos\theta$ to show that $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\sin^{-1} x$. Thus, X and Y have the same cdf and are called **arcsine random variables**. The corresponding density is $f_X(x) = (1/\pi)/\sqrt{1-x^2}$ for $-1 < x < 1$.

34. Find the cdf and density of $Y = X(X+2)$ if X is uniformly distributed on $[-3, 1]$.

35. Let g be as in Example 4.10 and 4.11. Find the cdf and density of $Y = g(X)$ if

    (a) $X \sim \text{uniform}[-1, 1]$;

    (b) $X \sim \text{uniform}[-1, 2]$;

    (c) $X \sim \text{uniform}[-2, 3]$;

    (d) $X \sim \exp(\lambda)$.

36. Let

$$g(x) := \begin{cases} 0, & |x| < 1, \\ |x| - 1, & 1 \le |x| \le 2, \\ 1, & |x| > 2. \end{cases}$$

    Find the cdf and density of $Y = g(X)$ if

    (a) $X \sim \text{uniform}[-1, 1]$;

    (b) $X \sim \text{uniform}[-2, 2]$;

    (c) $X \sim \text{uniform}[-3, 3]$;

    (d) $X \sim \text{Laplace}(\lambda)$.

37. Let

$$g(x) := \begin{cases} \Leftrightarrow x \Leftrightarrow 2, & x < \Leftrightarrow 1, \\ \Leftrightarrow x^2, & \Leftrightarrow 1 \le x < 0, \\ x^3, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

Find the cdf and density of $Y = g(X)$ if

(a) $X \sim \text{uniform}[\Leftrightarrow 3, 2]$;

(b) $X \sim \text{uniform}[\Leftrightarrow 3, 1]$;

(c) $X \sim \text{uniform}[\Leftrightarrow 1, 1]$.

38. Consider the function g given by

$$g(x) = \begin{cases} x^2 \Leftrightarrow 1, & x < 0, \\ x \Leftrightarrow 1, & 0 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

If X is uniform$[\Leftrightarrow 3, 3]$, find the cdf and density of $Y = g(X)$.

39. Let X be a uniformly distributed random variable on the interval $[\Leftrightarrow 3, 1]$. Let $Y = g(X)$, where

$$g(x) = \begin{cases} 0, & x < \Leftrightarrow 2, \\ x + 2, & \Leftrightarrow 2 \le x < \Leftrightarrow 1, \\ x^2, & \Leftrightarrow 1 \le x < 0, \\ \sqrt{x}, & x \ge 0. \end{cases}$$

Find the cdf and density of Y.

40. Let $X \sim \text{uniform}[\Leftrightarrow 6, 0]$, and suppose that $Y = g(X)$, where

$$g(x) = \begin{cases} |x| \Leftrightarrow 1, & 1 \le x < 2, \\ 1 \Leftrightarrow \sqrt{|x| \Leftrightarrow 2}, & |x| \ge 2, \\ 0, & \text{otherwise.} \end{cases}$$

Find the cdf and density of Y.

41. Let $X \sim \text{uniform}[\Leftrightarrow 2, 1]$, and suppose that $Y = g(X)$, where

$$g(x) = \begin{cases} x + 2, & \Leftrightarrow 2 \le x < \Leftrightarrow 1, \\ \dfrac{2x^2}{1 + x^2}, & \Leftrightarrow 1 \le x < 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find the cdf and density of Y.

*42. For $x \ge 0$, let $g(x)$ denote the fractional part of x. For example, $g(5.649) = 0.649$, and $g(0.123) = 0.123$. Find the cdf and density of $Y = g(X)$ if

(a) $X \sim \exp(1)$;

(b) $X \sim \text{uniform}[0,1)$;

(c) $X \sim \text{uniform}[v, v+1)$, where $v = m + \delta$ for some integer $m \geq 0$ and some $0 < \delta < 1$.

Problems §4.6: Properties of Cdfs

*43. From your solution of Problem 3(b) in Chapter 3, you can see that if $X \sim \exp(\lambda)$, then $\wp(X > t + \Delta t | X > t) = \wp(X > \Delta t)$. Now prove the converse; i.e., show that if Y is a nonnegative random variable such that $\wp(Y > t + \Delta t | Y > t) = \wp(Y > \Delta t)$, then $Y \sim \exp(\lambda)$, where $\lambda = \Leftrightarrow \ln[1 \Leftrightarrow F_Y(1)]$, assuming that $\wp(Y > t) > 0$ for all $t \geq 0$. *Hints:* Put $h(t) := \ln \wp(Y > t)$, which is a right-continuous function of t (Why?). Show that $h(t + \Delta t) = h(t) + h(\Delta t)$ for all $t, \Delta t \geq 0$.

Problems §4.7: The Central Limit Theorem

44. Packet transmission times on a certain Internet link are i.i.d. with mean m and variance σ^2. Suppose n packets are transmitted. Then the total expected transmission time for n packets is nm. Use the central limit theorem to approximate the probability that the total transmission time for the n packets exceeds twice the expected transmission time.

45. To combat noise in a digital communication channel with bit-error probability p, the use of an error-correcting code is proposed. Suppose that the code allows correct decoding of a received binary codeword if the fraction of bits in error is less than or equal to t. Use the central limit theorem to approximate the probability that a received word cannot be reliably decoded.

46. Let $X_i = \pm 1$ with equal probability. Then the X_i are zero mean and have unit variance. Put

$$Y_n \;=\; \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}.$$

Derive the central limit theorem for this case; i.e., show that $\varphi_{Y_n}(\nu) \to e^{-\nu^2/2}$. *Hint:* Use the Taylor series approximation $\cos(\xi) \approx 1 \Leftrightarrow \xi^2/2$.

CHAPTER 5

Multiple Random Variables

The main focus of this chapter is the study of multiple continuous random variables that are not independent. In particular, conditional probability and conditional expectation along with corresponding laws of total probability and substitution are studied. These tools are used to compute probabilities involving the output of systems with multiple random inputs.

In Section 5.1 we introduce the concept of joint cumulative distribution function for a pair of random variables. Marginal cdfs are also defined. In Section 5.2 we introduce pairs of jointly continuous random variables, joint densities, and marginal densities. Conditional densities, independence and expectation are also discussed in the context of jointly continuous random variables. In Section 5.3 conditional probability and expectation are defined for jointly continuous random variables. In Section 5.4 the bivariate normal density is introduced, and some of its properties are illustrated via examples. In Section 5.5 most of the bivariate concepts are extended to the case of n random variables.

5.1. Joint and Marginal Probabilities

Suppose we have a pair of random variables, say X and Y, for which we are able to compute $\wp((X,Y) \in A)$ for arbitrary[1] sets $A \subset \mathbb{R}^2$. For example, suppose X and Y are the horizontal and vertical coordinates of a dart striking a target. We might be interested in the probability that the dart lands within two units of the center. Then we would take A to be the disk of radius two centered at the origin, i.e., $A = \{(x,y) : x^2 + y^2 \le 4\}$. Now, even though we can write down a formula for $\wp((X,Y) \in A)$, we may be interested in finding potentially simpler expressions for things like $\wp(X \in B, Y \in C)$, $\wp(X \in B)$, and $\wp(Y \in C)$, etc. To this end, the notion of a product set is helpful.

Product Sets and Marginal Probabilities

The **Cartesian product** of two univariate sets B and C is defined by

$$B \times C \;:=\; \{(x,y) : x \in B \text{ and } y \in C\}.$$

In other words,

$$(x,y) \in B \times C \;\Leftrightarrow\; x \in B \text{ and } y \in C.$$

For example, if $B = [1,3]$ and $C = [0.5, 3.5]$, then $B \times C$ is the rectangle

$$[1,3] \times [0.5, 3.5] \;=\; \{(x,y) : 1 \le x \le 3 \text{ and } 0.5 \le y \le 3.5\},$$

which is illustrated in Figure 5.1(a). In general, if B and C are intervals, then $B \times C$ is a rectangle or square. If one of the sets is an interval and the other

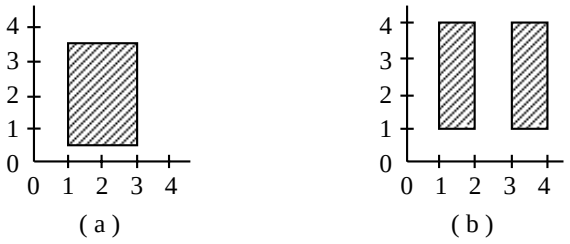

Figure 5.1. The Cartesian products (a) $[1,3] \times [0.5,3.5]$ and (b) $\big([1,2] \cup [3,4]\big) \times [1,4]$.

is a singleton, then the product set degenerates to a line segment in the plane. A more complicated example is shown in Figure 5.1(b), which illustrates the product $\big([1,2] \cup [3,4]\big) \times [1,4]$. Figure 5.1(b) also illustrates the general result that $\times$ distributes over $\cup$; i.e., $(B_1 \cup B_2) \times C = (B_1 \times C) \cup (B_2 \times C)$.

Using the notion of product set,

$$\begin{aligned} \{X \in B, Y \in C\} &= \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\} \\ &= \{\omega \in \Omega : (X(\omega), Y(\omega)) \in B \times C\}, \end{aligned}$$

for which we use the shorthand

$$\{(X,Y) \in B \times C\}.$$

We can therefore write

$$\wp(X \in B, Y \in C) = \wp((X,Y) \in B \times C).$$

The preceding expression allows us to obtain the **marginal probability** $\wp(X \in B)$ as follows. First, for any event F, we have $F \subset \Omega$, and therefore, $F = F \cap \Omega$. Second, Y is assumed to be a real-valued random variable, i.e., $Y(\omega) \in \mathbb{R}$ for all ω. Thus, $\{Y \in \mathbb{R}\} = \Omega$. Now write

$$\begin{aligned} \wp(X \in B) &= \wp(\{X \in B\} \cap \Omega) \\ &= \wp(\{X \in B\} \cap \{Y \in \mathbb{R}\}) \\ &= \wp(X \in B, Y \in \mathbb{R}) \\ &= \wp((X,Y) \in B \times \mathbb{R}). \end{aligned}$$

Similarly,

$$\wp(Y \in C) = \wp((X,Y) \in \mathbb{R} \times C).$$

If X and Y are independent random variables, then

$$\begin{aligned} \wp((X,Y) \in B \times C) &= \wp(X \in B, Y \in C) \\ &= \wp(X \in B)\, \wp(Y \in C). \end{aligned}$$

Joint and Marginal Cumulative Distributions

The **joint cumulative distribution function** of X and Y is defined by

$$\begin{aligned} F_{XY}(x,y) &:= \wp(X \le x, Y \le y) \\ &= \wp((X,Y) \in (\Leftrightarrow\infty, x] \times (\Leftrightarrow\infty, y]). \end{aligned}$$

In Chapter 4, we saw that $\wp(X \in B)$ could be computed if we knew the cumulative distribution function $F_X(x)$ for all x. Similarly, it turns out that we can compute $\wp((X,Y) \in A)$ if we know $F_{XY}(x,y)$ for all x, y. For example, it is shown in Problems 1 and 2 that

$$\wp((X,Y) \in (a,b] \times (c,d]) \;=\; \wp(a < X \le b, c < Y \le d)$$

is given by

$$F_{XY}(b,d) \Leftrightarrow F_{XY}(a,d) \Leftrightarrow F_{XY}(b,c) + F_{XY}(a,c).$$

In other words, the probability that (X,Y) lies in a rectangle can be computed in terms of the cdf values at the corners.

If X and Y are independent random variables, we have the simplifications

$$\begin{aligned} F_{XY}(x,y) &= \wp(X \le x, Y \le y) \\ &= \wp(X \le x)\,\wp(Y \le y) \\ &= F_X(x) F_Y(y), \end{aligned}$$

and

$$\wp(a < X \le b, c < Y \le d) = \wp(a < X \le b)\,\wp(c < Y \le d),$$

which is equal to $[F_X(b) \Leftrightarrow F_X(a)][F_Y(d) \Leftrightarrow F_Y(c)]$.

We return now to the general case (X and Y not necessarily independent). It is possible to obtain the **marginal cumulative distributions** F_X and F_Y directly from F_{XY} by setting the unwanted variable to ∞. More precisely, it can be shown that[2]

$$F_X(x) \;=\; \lim_{y\to\infty} F_{XY}(x,y) \;=:\; F_{XY}(x,\infty), \tag{5.1}$$

and

$$F_Y(y) \;=\; \lim_{x\to\infty} F_{XY}(x,y) \;=:\; F_{XY}(\infty,y).$$

***Example* 5.1.** If

$$F_{XY}(x,y) \;=\; \begin{cases} \dfrac{y + e^{-x(y+1)}}{y+1} \Leftrightarrow e^{-x}, & x, y > 0, \\ 0, & \text{otherwise}, \end{cases}$$

find both of the marginal cumulative distribution functions, $F_X(x)$ and $F_Y(y)$.

Solution. For $x, y > 0$,

$$F_{XY}(x,y) \;=\; \frac{y}{y+1} + \frac{1}{y+1}\cdot e^{-x(y+1)} \Leftrightarrow e^{-x}.$$

Hence, for $x > 0$,

$$\lim_{y\to\infty} F_{XY}(x,y) \;=\; 1 + 0\cdot 0 \Leftrightarrow e^{-x} \;=\; 1 \Leftrightarrow e^{-x}.$$

For $x \le 0$, $F_{XY}(x,y) = 0$ for all y. So, for $x \le 0$, $\lim_{y\to\infty} F_{XY}(x,y) = 0$. The complete formula for the marginal cdf of X is

$$F_X(x) \;=\; \begin{cases} 1 \Leftrightarrow e^{-x}, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{5.2}$$

Next, for $y > 0$,

$$\lim_{x\to\infty} F_{XY}(x,y) \;=\; \frac{y}{y+1} + \frac{1}{y+1}\cdot 0 \Leftrightarrow 0 \;=\; \frac{y}{y+1}.$$

We then see that the marginal cdf of Y is

$$F_Y(y) \;=\; \begin{cases} y/(y+1), & y > 0, \\ 0, & y \le 0. \end{cases} \tag{5.3}$$

Note that since $F_{XY}(x,y) \neq F_X(x)\,F_Y(y)$ for all x, y, X and Y are *not* independent,

---

***Example* 5.2.** Express $\wp(X \le x \text{ or } Y \le y)$ using cdfs.

Solution. The desired probability is that of the union, $\wp(A \cup B)$, where $A := \{X \le x\}$ and $B := \{Y \le y\}$. Applying the inclusion-exclusion formula (1.1) yields

$$\begin{aligned} \wp(A \cup B) &= \wp(A) + \wp(B) \Leftrightarrow \wp(A \cap B) \\ &= \wp(X \le x) + \wp(Y \le y) \Leftrightarrow \wp(X \le x, Y \le y) \\ &= F_X(x) + F_Y(y) \Leftrightarrow F_{XY}(x,y). \end{aligned}$$

5.2. Jointly Continuous Random Variables

In analogy with the univariate case, we say that two random variables X and Y are **jointly continuous** with **joint density** $f_{XY}(x,y)$ if

$$\wp((X,Y) \in A) \;=\; \iint_A f_{XY}(x,y)\,dx\,dy$$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_A(x,y) f_{XY}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} I_A(x,y) f_{XY}(x,y)\,dy\right)dx \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} I_A(x,y) f_{XY}(x,y)\,dx\right)dy \end{aligned}$$

for some nonnegative function f_{XY} that integrates to one; i.e.,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy \;=\; 1.$$

***Example* 5.3.** Show that

$$f_{XY}(x,y) \;=\; \tfrac{1}{2\pi} e^{-(2x^2-2xy+y^2)/2}$$

is a valid joint probability density.

Solution. Since $f_{XY}(x,y)$ is nonnegative, all we have to do is show that it integrates to one. By completing the square in the exponent, we obtain

$$f_{XY}(x,y) \;=\; \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}} \cdot \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

This factorization allows us to write the double integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{-(2x^2-2xy+y^2)/2}}{2\pi}\,dx\,dy$$

as the iterated integral

$$\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\left(\int_{-\infty}^{\infty} \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}\,dy\right)dx.$$

The inner integral, as a function of y, is a normal density with mean x and variance one. Hence, the inner integral is one. But this leaves only the outer integral, whose integrand is an $N(0,1)$ density, which also integrates to one.

Marginal Densities

If A is a product set, say $A = B \times C$, then $I_A(x,y) = I_B(x)\, I_C(y)$, and so

$$\begin{aligned} \mathcal{P}(X \in B, Y \in C) &= \mathcal{P}((X,Y) \in B \times C) \\ &= \int_B \left(\int_C f_{XY}(x,y)\,dy\right)dx \qquad (5.4) \\ &= \int_C \left(\int_B f_{XY}(x,y)\,dx\right)dy. \end{aligned}$$

At this point we would like to substitute $B = (\Leftrightarrow\infty, x]$ and $C = (\Leftrightarrow\infty, y]$ in order to obtain expressions for $F_{XY}(x,y)$. However, the preceding integrals already use x and y for the variables of integration. To avoid confusion, we must first replace the variables of integration. We change x to t and y to τ. We then find that

$$F_{XY}(x,y) \;=\; \int_{-\infty}^{x}\left(\int_{-\infty}^{y} f_{XY}(t,\tau)\,d\tau\right)dt,$$

or, equivalently,

$$F_{XY}(x,y) \;=\; \int_{-\infty}^{y}\left(\int_{-\infty}^{x} f_{XY}(t,\tau)\,dt\right)d\tau.$$

It then follows that

$$\frac{\partial^2}{\partial y\partial x}F_{XY}(x,y) \;=\; f_{XY}(x,y) \quad\text{and}\quad \frac{\partial^2}{\partial x\partial y}F_{XY}(x,y) \;=\; f_{XY}(x,y).$$

***Example* 5.4.** Let

$$F_{XY}(x,y) \;=\; \begin{cases} \dfrac{y+e^{-x(y+1)}}{y+1} \Leftrightarrow e^{-x}, & x,y>0,\\ 0, & \text{otherwise,}\end{cases}$$

as in Example 5.1. Find the joint density f_{XY}.

Solution. For $x, y > 0$,

$$\frac{\partial}{\partial x}F_{XY}(x,y) \;=\; e^{-x} \Leftrightarrow e^{-x(y+1)},$$

and

$$\frac{\partial^2}{\partial y\partial x}F_{XY}(x,y) \;=\; xe^{-x(y+1)}.$$

Thus,

$$f_{XY}(x,y) \;=\; \begin{cases} xe^{-x(y+1)}, & x,y>0,\\ 0, & \text{otherwise.}\end{cases}$$

We now show that if X and Y are jointly continuous, then X and Y are individually continuous with marginal densities obtained as follows. Taking $C = \mathbb{R}$ in (5.4), we obtain

$$\wp(X\in B) \;=\; \wp((X,Y)\in B\times\mathbb{R}) \;=\; \int_B\left(\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy\right)dx,$$

which implies the **marginal density** of X is

$$f_X(x) \;=\; \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy. \tag{5.5}$$

Similarly,

$$\wp(Y \in C) \;=\; \wp((X,Y) \in \mathbb{R} \times C) \;=\; \int_C \left(\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx \right) dy,$$

and

$$f_Y(y) \;=\; \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx.$$

Thus, to obtain the marginal densities, integrate out the unwanted variable.

***Example* 5.5.** Using the joint density f_{XY} obtained in Example 5.4, find the marginal densities f_X and f_Y by integrating out the unneeded variable. To check your answer, also compute the marginal densities by differentiating the marginal cdfs obtained in Example 5.1.

Solution. We first compute $f_X(x)$. To begin, observe that for $x \le 0$, $f_{XY}(x,y) = 0$. Hence, for $x \le 0$, the integral in (5.5) is zero. Now suppose $x > 0$. Since $f_{XY}(x,y) = 0$ whenever $y \le 0$, the lower limit of integration in (5.5) can be changed to zero. For $x > 0$, it remains to compute

$$\begin{aligned} \int_0^{\infty} f_{XY}(x,y)\,dy &= xe^{-x} \int_0^{\infty} e^{-xy}\,dy \\ &= e^{-x}. \end{aligned}$$

Hence,

$$f_X(x) \;=\; \begin{cases} e^{-x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

and we see that X is exponentially distributed with parameter $\lambda = 1$. Note that the same answer can be obtained by differentiating the formula for $F_X(x)$ in (5.2).

We now turn to the calculation of $f_Y(y)$. Arguing as above, we have $f_Y(y) = 0$ for $y \le 0$, and $f_Y(y) = \int_0^{\infty} f_{XY}(x,y)\,dx$ for $y > 0$. Write this integral as

$$\int_0^{\infty} f_{XY}(x,y)\,dx \;=\; \frac{1}{y+1} \int_0^{\infty} x \cdot (y+1) e^{-(y+1)x}\,dx. \tag{5.6}$$

If we put $\lambda = y + 1$, then the integral on the right has the form

$$\int_0^{\infty} x \cdot \lambda e^{-\lambda x}\,dx,$$

which is the mean of an exponential random variable with parameter λ. This integral is equal to $1/\lambda = 1/(y+1)$, and so the right-hand side of (5.6) is equal to $1/(y+1)^2$. We conclude that

$$f_Y(y) \;=\; \begin{cases} 1/(y+1)^2, & y > 0, \\ 0, & y \le 0. \end{cases}$$

Note that the same answer can be obtained by differentiating the formula for $F_Y(y)$ in (5.3).

Specifying Joint Densities

Suppose X and Y are jointly continuous with joint density f_{XY}. Define

$$f_{Y|X}(y|x) \;:=\; \frac{f_{XY}(x,y)}{f_X(x)}. \tag{5.7}$$

For reasons that will become clear later, we call $f_{Y|X}$ the **conditional density** of Y given X. For the moment, simply note that as a function of y, $f_{Y|X}(y|x)$ is nonnegative and integrates to one, since

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)\,dy \;=\; \frac{\displaystyle\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy}{f_X(x)} \;=\; \frac{f_X(x)}{f_X(x)} \;=\; 1.$$

Now rewrite (5.7) as

$$f_{XY}(x,y) \;=\; f_{Y|X}(y|x)\,f_X(x).$$

This suggests an easy way to specify a joint density. First, pick any one-dimensional density, say $f_X(x)$. Then $f_X(x)$ is nonnegative and integrates to one. Next, for each x, let $f_{Y|X}(y|x)$ be any density in the variable y. For different values of x, $f_{Y|X}(y|x)$ could be very different densities as a function of y; the only important thing is that for each x, $f_{Y|X}(y|x)$ as a function of y should be nonnegative and integrate to one with respect to y. Now if we *define* $f_{XY}(x,y) := f_{Y|X}(y|x)\,f_X(x)$, we will have a nonnegative function of (x,y) whose double integral is one. In practice, and in the problems at the end of the chapter, this is how we usually specify joint densities.

***Example* 5.6.** Let $X \sim N(0,1)$, and suppose that the conditional density of Y given $X = x$ is $N(x,1)$. Find the joint density $f_{XY}(x,y)$.

Solution. First note that

$$f_X(x) \;=\; \frac{e^{-x^2/2}}{\sqrt{2\pi}} \quad\text{and}\quad f_{Y|X}(y|x) \;=\; \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}.$$

Then write

$$f_{XY}(x,y) \;=\; f_{Y|X}(y|x)\,f_X(x) \;=\; \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}} \cdot \frac{e^{-x^2/2}}{\sqrt{2\pi}},$$

which simplifies to

$$f_{XY}(x,y) \;=\; \frac{\exp[-(2x^2 - 2xy + y^2)/2]}{2\pi}.$$

By interchanging the roles of X and Y, we have

$$f_{X|Y}(x|y) \;:=\; \frac{f_{XY}(x,y)}{f_Y(y)},$$

and we can define $f_{XY}(x,y) = f_{X|Y}(x|y)\,f_Y(y)$ if we specify a density $f_Y(y)$ and a function $f_{X|Y}(x|y)$ that is a density in x for each fixed y.

Independence

We now consider the joint density of jointly continuous *independent* random variables. As noted in Section 5.1, if X and Y are independent, then $F_{XY}(x,y) = F_X(x)\,F_Y(y)$ for all x and y. If X and Y are also jointly continuous, then by taking second-order mixed partial derivatives, we find

$$\frac{\partial^2}{\partial y \partial x} F_X(x)\,F_Y(y) \;=\; f_X(x)\,f_Y(y).$$

In other words, if X and Y are jointly continuous and independent, then the joint density is the product of the marginal densities. Using (5.4), it is easy to see that the converse is also true. If $f_{XY}(x,y) = f_X(x)\,f_Y(y)$, (5.4) implies

$$\begin{aligned}
\wp(X \in B, Y \in C) &= \int_B \biggl(\int_C f_{XY}(x,y)\,dy \biggr) dx \\
&= \int_B \biggl(\int_C f_X(x)\,f_Y(y)\,dy \biggr) dx \\
&= \int_B f_X(x) \biggl(\int_C f_Y(y)\,dy \biggr) dx \\
&= \biggl(\int_B f_X(x)\,dx \biggr) \wp(Y \in C) \\
&= \wp(X \in B)\,\wp(Y \in C).
\end{aligned}$$

Expectation

If X and Y are jointly continuous with joint density f_{XY}, then the methods of Section 3.2 can easily be used to show that

$$\mathsf{E}[g(X,Y)] \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\,dx\,dy.$$

For arbitrary random variables X and Y, their **bivariate characteristic function** is defined by

$$\varphi_{XY}(\nu_1,\nu_2) \;:=\; \mathsf{E}[e^{j(\nu_1 X + \nu_2 Y)}].$$

If X and Y have joint density f_{XY}, then

$$\varphi_{XY}(\nu_1,\nu_2) \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) e^{j(\nu_1 x + \nu_2 y)}\,dx\,dy,$$

which is simply the **bivariate Fourier transform** of f_{XY}. By the inversion formula,

$$f_{XY}(x,y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi_{XY}(\nu_1,\nu_2) e^{-j(\nu_1 x + \nu_2 y)}\,d\nu_1\,d\nu_2.$$

Now suppose that X and Y are independent. Then

$$\varphi_{XY}(\nu_1,\nu_2) \;=\; \mathsf{E}[e^{j(\nu_1 X + \nu_2 Y)}] \;=\; \mathsf{E}[e^{j\nu_1 X}]\,\mathsf{E}[e^{j\nu_2 Y}] \;=\; \varphi_X(\nu_1)\,\varphi_Y(\nu_2).$$

In other words, if X and Y are independent, then their joint characteristic function factors. The converse is also true; i.e., if the joint characteristic function factors, then X and Y are independent. The general proof is complicated, but if X and Y are jointly continuous, it is easy using the Fourier inversion formula.

⋆ Continuous Random Variables That Are not Jointly Continuous

Let $\Theta \sim \text{uniform}[\Leftrightarrow\pi, \pi]$, and put $X := \cos\Theta$ and $Y := \sin\Theta$. As shown in Problem 33 in Chapter 4, X and Y are both arcsine random variables, each having density $(1/\pi)/\sqrt{1 \Leftrightarrow x^2}$ for $\Leftrightarrow 1 < x < 1$.

Next, since $X^2 + Y^2 = 1$, the pair (X, Y) takes values only on the unit circle

$$C \;:=\; \{(x, y) : x^2 + y^2 = 1\}.$$

Thus, $\wp\big((X, Y) \in C\big) = 1$. On the other hand, if X and Y have a joint density f_{XY}, then

$$\wp\big((X, Y) \in C\big) \;=\; \iint\limits_C f_{XY}(x, y)\, dx\, dy \;=\; 0$$

because a double integral over a set of zero area must be zero. So, if X and Y had a joint density, this would imply that $1 = 0$. Since this is not true, there can be no joint density.

5.3. Conditional Probability and Expectation

If X is a continuous random variable, then its cdf $F_X(x) := \wp(X \le x) = \int_{-\infty}^{x} f_X(t)\, dt$ is a continuous function of x. It follows from the properties of cdfs in Section 4.6 that $\wp(X = x) = 0$ for all x. Hence, we cannot define $\wp(Y \in C | X = x)$ by $\wp(X = x, Y \in C)/\wp(X = x)$ since this requires division by zero. Similar problems arise with conditional expectation. How should we define conditional probability and expectation in this case? Recall from Section 2.5 that if X and Y are both discrete random variables, then the following law of total probability holds,

$$\mathsf{E}[g(X, Y)] \;=\; \sum_i \mathsf{E}[g(X, Y) | X = x_i]\, p_X(x_i).$$

If X and Y are jointly continuous, however we define $\mathsf{E}[g(X, Y) | X = x]$, we certainly want the analogous formula

$$\mathsf{E}[g(X, Y)] \;=\; \int_{-\infty}^{\infty} \mathsf{E}[g(X, Y) | X = x] f_X(x)\, dx \tag{5.8}$$

to hold. To discover how $\mathsf{E}[g(X, Y) | X = x]$ should be defined, write

$$\begin{aligned} \mathsf{E}[g(X, Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dy \right) dx \end{aligned}$$

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y) \frac{f_{XY}(x,y)}{f_X(x)}\, dy \right) f_X(x)\, dx$$
$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y|x)\, dy \right) f_X(x)\, dx.$$

It follows that if

$$\mathsf{E}[g(X,Y)|X=x] := \int_{-\infty}^{\infty} g(x,y)\, f_{Y|X}(y|x)\, dy, \tag{5.9}$$

then (5.8) holds.

Now observe that if g in (5.9) is a function of y alone, then

$$\mathsf{E}[g(Y)|X=x] = \int_{-\infty}^{\infty} g(y)\, f_{Y|X}(y|x)\, dy. \tag{5.10}$$

Returning now to the case $g(x,y)$, we see that (5.10) implies that for any $\tilde{x}$,

$$\mathsf{E}[g(\tilde{x},Y)|X=x] = \int_{-\infty}^{\infty} g(\tilde{x},y)\, f_{Y|X}(y|x)\, dy.$$

Taking $\tilde{x} = x$, we obtain

$$\mathsf{E}[g(x,Y)|X=x] = \int_{-\infty}^{\infty} g(x,y)\, f_{Y|X}(y|x)\, dy.$$

Comparing this with (5.9) shows that

$$\mathsf{E}[g(X,Y)|X=x] = \mathsf{E}[g(x,Y)|X=x].$$

In other words, the substitution law holds. Another important point to note is that if X and Y are independent, then

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_X(x)\, f_Y(y)}{f_X(x)} = f_Y(y).$$

In this case, (5.10) becomes $\mathsf{E}[g(Y)|X=x] = \mathsf{E}[g(Y)]$. In other words, we can "drop the conditioning."

***Example* 5.7.** Let $X \sim \exp(1)$, and suppose that given $X = x$, Y is conditionally normal with $f_{Y|X}(\cdot|x) \sim N(0, x^2)$. Evaluate $\mathsf{E}[Y^2]$ and $\mathsf{E}[Y^2X^3]$.

Solution. We use the law of total probability for expectation. We begin with

$$\mathsf{E}[Y^2] = \int_{-\infty}^{\infty} \mathsf{E}[Y^2|X=x] f_X(x)\, dx.$$

Since $f_{Y|X}(y|x) = e^{-(y/x)^2/2}/(\sqrt{2\pi}\,x)$, we see that

$$\mathsf{E}[Y^2|X=x] \;=\; \int_{-\infty}^{\infty} y^2\,\frac{e^{-(y/x)^2/2}}{\sqrt{2\pi}\,x}\,dy$$

is the second moment of a zero-mean Gaussian density with variance x^2. Hence, $\mathsf{E}[Y^2|X=x] = x^2$, and

$$\mathsf{E}[Y^2] \;=\; \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \;=\; \mathsf{E}[X^2].$$

Since $X \sim \exp(1)$, $\mathsf{E}[X^2] = 2$ by Example 3.9.

To compute $\mathsf{E}[Y^2X^3]$, we proceed similarly. Write

$$\begin{aligned}
\mathsf{E}[Y^2X^3] &= \int_{-\infty}^{\infty} \mathsf{E}[Y^2X^3|X=x]f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} \mathsf{E}[Y^2x^3|X=x]f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} x^3\mathsf{E}[Y^2|X=x]f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} x^3x^2 f_X(x)\,dx \\
&= \mathsf{E}[X^5] \\
&= 5!, \quad \text{by Example 3.9.}
\end{aligned}$$

For conditional probability, if we put

$$\wp(Y \in C|X=x) \;:=\; \int_C f_{Y|X}(y|x)\,dy,$$

then it is easy to show that the law of total probability

$$\wp(Y \in C) \;=\; \int_{-\infty}^{\infty} \wp(Y \in C|X=x)\,f_X(x)\,dx$$

holds. It can also be shown[3] that the substitution law holds for conditional probabilities involving X and Y conditioned on X. Also, if X and Y are independent, then $\wp(Y \in C|X=x) = \wp(Y \in C)$; i.e., we can "drop the conditioning."

***Example* 5.8** (Signal in Additive Noise). A random, continuous-valued signal X is transmitted over a channel subject to additive, continuous-valued noise Y. The received signal is $Z = X + Y$. Find the cdf and density of Z if X and Y are jointly continuous random variables with joint density f_{XY}.

Solution. Since we are not assuming that X and Y are independent, the characteristic-function method of Example 3.14 does not work here. Instead, we use the laws of total probability and substitution. Write

$$\begin{aligned} F_Z(z) &= \wp(Z \le z) \\ &= \int_{-\infty}^{\infty} \wp(Z \le z | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \wp(X + Y \le z | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \wp(X + y \le z | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \wp(X \le z \Leftrightarrow y | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} F_{X|Y}(z \Leftrightarrow y | y)\, f_Y(y)\, dy. \end{aligned}$$

By differentiating with respect to z,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X|Y}(z \Leftrightarrow y | y)\, f_Y(y)\, dy.$$

In particular, note that if X and Y are independent, we can drop the conditioning and recover the convolution result stated following Example 3.14,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z \Leftrightarrow y)\, f_Y(y)\, dy.$$

***Example* 5.9** (Signal in Multiplicative Noise). A random, continuous-valued signal X is transmitted over a channel subject to multiplicative, continuous-valued noise Y. The received signal is $Z = XY$. Find the cdf and density of Z if X and Y are jointly continuous random variables with joint density f_{XY}.

Solution. We proceed as in the previous example. Write

$$\begin{aligned} F_Z(z) &= \wp(Z \le z) \\ &= \int_{-\infty}^{\infty} \wp(Z \le z | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \wp(XY \le z | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \wp(Xy \le z | Y = y)\, f_Y(y)\, dy. \end{aligned}$$

At this point we have a problem when we attempt to divide through by y. If y is negative, we have to reverse the inequality sign. Otherwise, we do not have

to reverse the inequality. The solution to this difficulty is to break up the range of integration. Write

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{0} \wp(Xy \le z | Y = y)\, f_Y(y)\, dy \\ &\quad + \int_{0}^{\infty} \wp(Xy \le z | Y = y)\, f_Y(y)\, dy. \end{aligned}$$

Now we can divide by y separately in each integral. Thus,

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{0} \wp(X \ge z/y | Y = y)\, f_Y(y)\, dy \\ &\quad + \int_{0}^{\infty} \wp(X \le z/y | Y = y)\, f_Y(y)\, dy \\ &= \int_{-\infty}^{0} \left[1 - F_{X|Y}\left(\tfrac{z}{y}|y\right)\right] f_Y(y)\, dy \\ &\quad + \int_{0}^{\infty} F_{X|Y}\left(\tfrac{z}{y}|y\right) f_Y(y)\, dy. \end{aligned}$$

Differentiating with respect to z yields

$$f_Z(z) = -\int_{-\infty}^{0} f_{X|Y}\left(\tfrac{z}{y}|y\right)\tfrac{1}{y}\, f_Y(y)\, dy + \int_{0}^{\infty} f_{X|Y}\left(\tfrac{z}{y}|y\right)\tfrac{1}{y}\, f_Y(y)\, dy.$$

Now observe that in the first integral, the range of integration implies that y is always negative. For such y, $-y = |y|$. In the second integral, y is always positive, and so $y = |y|$. Thus,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{0} f_{X|Y}\left(\tfrac{z}{y}|y\right)\tfrac{1}{|y|}\, f_Y(y)\, dy + \int_{0}^{\infty} f_{X|Y}\left(\tfrac{z}{y}|y\right)\tfrac{1}{|y|}\, f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} f_{X|Y}\left(\tfrac{z}{y}|y\right)\tfrac{1}{|y|}\, f_Y(y)\, dy. \end{aligned}$$

5.4. The Bivariate Normal

The **bivariate Gaussian** or **bivariate normal** density is a generalization of the univariate $N(m, \sigma^2)$ density. Recall that the standard $N(0,1)$ density is given by $\varphi(x) := \exp(-x^2/2)/\sqrt{2\pi}$. The general $N(m, \sigma^2)$ density can be written in terms of φ as

$$\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right] = \frac{1}{\sigma} \cdot \varphi\left(\frac{x-m}{\sigma}\right).$$

In order to define the general bivariate Gaussian density, it is convenient to define a standard bivariate density first. So, for $|\rho| < 1$, put

$$\varphi_\rho(u,v) := \frac{\exp\left(\frac{-1}{2(1-\rho^2)}[u^2 - 2\rho uv + v^2]\right)}{2\pi\sqrt{1-\rho^2}}. \tag{5.11}$$

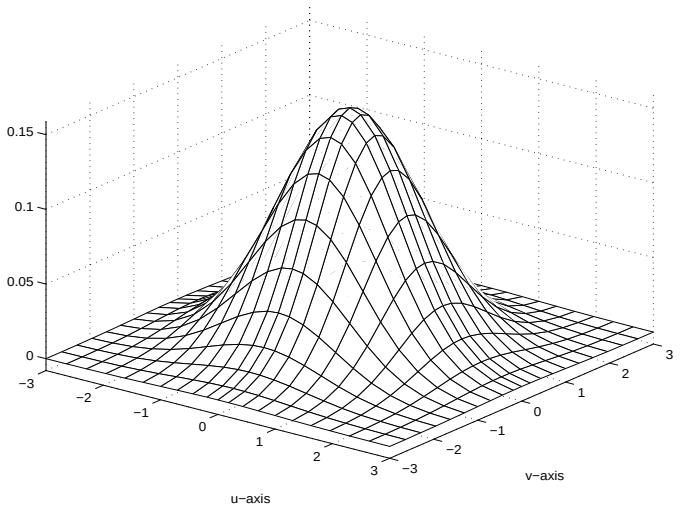

Figure 5.2. The Gaussian surface $\varphi_\rho(u,v)$ of (5.11) with $\rho = 0$.

For fixed ρ, this function of the two variables u and v defines a surface. The surface corresponding to $\rho = 0$ is shown in Figure 5.2. From the figure and from the formula (5.11), we see that φ_0 is circularly symmetric; i.e., for all (u,v) on a circle of radius r, in other words, for $u^2+v^2 = r^2$, $\varphi_0(u,v) = e^{-r^2/2}/2\pi$ does not depend on the particular values of u and v, but only on the radius of the circle on which they lie. We also point out that for $\rho = 0$, the formula (5.11) factors into the product of two univariate $N(0,1)$ densities, i.e., $\varphi_0(u,v) = \varphi(u)\,\varphi(v)$. For $\rho \neq 0$, φ_ρ does not factor. In other words, if U and V have joint density φ_ρ, then U and V are independent if and only if $\rho = 0$. A plot of φ_ρ for $\rho = \Leftrightarrow 0.85$ is shown in Figure 5.3. It turns out that now φ_ρ is constant on ellipses instead of circles. The axes of the ellipses are not parallel to the coordinate axes. Notice how the density is concentrated along the line $v = \Leftrightarrow u$. As $\rho \to \Leftrightarrow 1$, this concentration becomes more extreme. As $\rho \to +1$, the density concentrates around the line $v = u$. We now show that the density φ_ρ integrates to one. To do this, first observe that for all $|\rho| < 1$,

$$u^2 \Leftrightarrow 2\rho uv + v^2 \;=\; u^2(1 \Leftrightarrow \rho^2) + (v \Leftrightarrow \rho u)^2.$$

It follows that

$$\begin{aligned} \varphi_\rho(u,v) &= \frac{e^{-u^2/2}}{\sqrt{2\pi}} \cdot \frac{\exp\left(\frac{-1}{2(1-\rho^2)}[v \Leftrightarrow \rho u]^2\right)}{\sqrt{2\pi}\sqrt{1 \Leftrightarrow \rho^2}} \\ &= \varphi(u) \cdot \frac{1}{\sqrt{1 \Leftrightarrow \rho^2}}\,\varphi\left(\frac{v \Leftrightarrow \rho u}{\sqrt{1 \Leftrightarrow \rho^2}}\right). \end{aligned} \tag{5.12}$$

Observe that the right-hand factor as a function of v has the form of a univariate

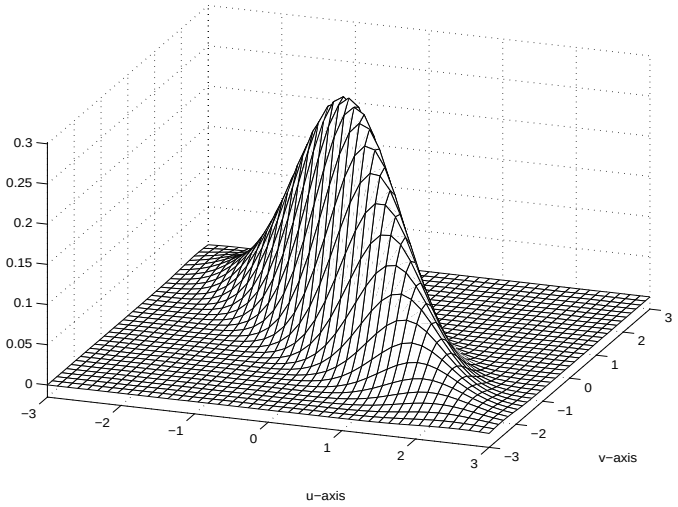

Figure 5.3. The Gaussian surface $\varphi_\rho(u,v)$ of (5.11) with $\rho = -0.85$.

normal density with mean ρu and variance $1 \Leftrightarrow \rho^2$. With φ_ρ factored as in (5.12), we can write $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\rho(u,v)\,du\,dv$ as the iterated integral

$$\int_{-\infty}^{\infty} \varphi(u)\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{1 \Leftrightarrow \rho^2}}\varphi\left(\frac{v \Leftrightarrow \rho u}{\sqrt{1 \Leftrightarrow \rho^2}}\right)dv\right]du.$$

As noted above, the inner integrand, as a function of v, is simply an $N(\rho u, 1 \Leftrightarrow \rho^2)$ density, and therefore integrates to one. Hence, the above iterated integral becomes $\int_{-\infty}^{\infty} \varphi(u)\,du = 1$.

We can now easily define the general bivariate Gaussian density with parameters m_X, m_Y, σ_X^2, σ_Y^2, and ρ by

$$f_{XY}(x,y) \;:=\; \frac{1}{\sigma_X\sigma_Y}\varphi_\rho\left(\frac{x \Leftrightarrow m_X}{\sigma_X}, \frac{y \Leftrightarrow m_Y}{\sigma_Y}\right).$$

More explicitly, this density is

$$\frac{\exp\left(\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-m_X}{\sigma_X}\right)^2 \Leftrightarrow 2\rho\left(\frac{x-m_X}{\sigma_X}\right)\left(\frac{y-m_Y}{\sigma_Y}\right) + \left(\frac{y-m_Y}{\sigma_Y}\right)^2\right]\right)}{2\pi\sigma_X\sigma_Y\sqrt{1 \Leftrightarrow \rho^2}}. \tag{5.13}$$

It can be shown that the marginals are $f_X \sim N(m_X, \sigma_X^2)$ and $f_Y \sim N(m_Y, \sigma_Y^2)$ (see Problems 25 and 28). The parameter ρ is called the **correlation coefficient**. From (5.13), we observe that X and Y are independent if and only if $\rho = 0$. A plot of f_{XY} with $m_X = m_Y = 0$, $\sigma_X = 1.5$, $\sigma_Y = 0.6$, and $\rho = 0$ is shown in Figure 5.4. Notice how the density is concentrated around the line

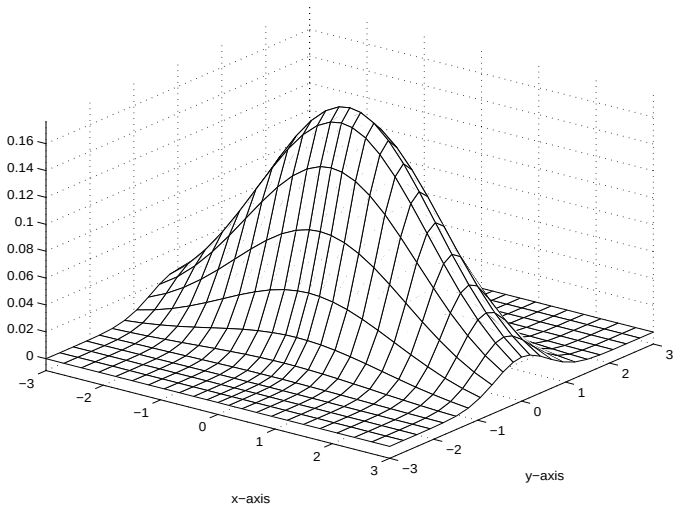

Figure 5.4. The bivariate normal density $f_{XY}(x,y)$ of (5.13) with $m_X = m_Y = 0$, $\sigma_X = 1.5$, $\sigma_Y = 0.6$, and $\rho = 0$.

$y = 0$. Also, f_{XY} is constant on ellipses of the form

$$\left(\frac{x}{\sigma_X}\right)^2 + \left(\frac{y}{\sigma_Y}\right)^2 \;=\; r^2.$$

To show that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1$ as well, use formula (5.12) for φ_ρ and proceed as above, integrating with respect to y first and then x. For the inner integral, make the change of variable $v = (y \Leftrightarrow m_Y)/\sigma_Y$, and in the remaining outer integral make the change of variable $u = (x \Leftrightarrow m_X)/\sigma_X$.

***Example* 5.10.** Let random variables U and V have the standard bivariate normal density φ_ρ in (5.11). Show that $\mathsf{E}[UV] = \rho$.

Solution. Using the factored form of φ_ρ in (5.12), write

$$\begin{aligned}\mathsf{E}[UV] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv\,\varphi_\rho(u,v)\,du\,dv \\ &= \int_{-\infty}^{\infty} u\,\varphi(u)\left[\int_{-\infty}^{\infty}\frac{v}{\sqrt{1\Leftrightarrow\rho^2}}\,\varphi\left(\frac{v\Leftrightarrow\rho u}{\sqrt{1\Leftrightarrow\rho^2}}\right)dv\right]du.\end{aligned}$$

The quantity in brackets has the form $\mathsf{E}[\hat{V}]$, where $\hat{V}$ is a univariate normal random variable with mean ρu and variance $1 \Leftrightarrow \rho^2$. Thus,

$$\begin{aligned}\mathsf{E}[UV] &= \int_{-\infty}^{\infty} u\,\varphi(u)[\rho u]\,du \\ &= \rho\int_{-\infty}^{\infty} u^2\,\varphi(u)\,du \\ &= \rho,\end{aligned}$$

since φ is the $N(0,1)$ density.

***Example* 5.11.** Let U and V have the standard bivariate normal density $f_{UV}(u,v) = \varphi_\rho(u,v)$ given in (5.11). Find the conditional densities $f_{V|U}$ and $f_{U|V}$.

Solution. It is shown in Problem 25 that f_U and f_V are both $N(0,1)$. Hence,

$$f_{V|U}(v|u) \;=\; \frac{f_{UV}(u,v)}{f_U(u)} \;=\; \frac{\varphi_\rho(u,v)}{\varphi(u)},$$

where φ is the $N(0,1)$ density. If we now substitute the factored form of $\varphi_\rho(u,v)$ given in (5.12), we obtain

$$f_{V|U}(v|u) \;=\; \frac{1}{\sqrt{1-\rho^2}}\,\varphi\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right);$$

i.e., $f_{V|U}(\cdot\,|u) \sim N(\rho u, 1-\rho^2)$. To compute $f_{U|V}$ we need the following alternative factorization of φ_ρ,

$$\varphi_\rho(u,v) \;=\; \frac{1}{\sqrt{1-\rho^2}}\,\varphi\left(\frac{u-\rho v}{\sqrt{1-\rho^2}}\right)\cdot\varphi(v).$$

It then follows that

$$f_{U|V}(u|v) \;=\; \frac{1}{\sqrt{1-\rho^2}}\,\varphi\left(\frac{u-\rho v}{\sqrt{1-\rho^2}}\right);$$

i.e., $f_{U|V}(\cdot\,|v) \sim N(\rho v, 1-\rho^2)$.

***Example* 5.12.** If U and V have standard joint normal density $\varphi_\rho(u,v)$, find $\mathsf{E}[V|U=u]$.

Solution. Recall that from Example 5.11, $f_{V|U}(\cdot|u) \sim N(\rho u, 1-\rho^2)$. Hence,

$$\mathsf{E}[V|U=u] \;=\; \int_{-\infty}^{\infty} v\, f_{V|U}(v|u)\,dv \;=\; \rho u.$$

5.5. *Multivariate Random Variables

Let $X := [X_1,\ldots,X_n]'$ (the prime denotes transpose) be a length-n column vector of random variables. We call X a **random vector**. We say that the variables $X_1,\ldots,X_n$ are **jointly continuous** with **joint density** $f_X = f_{X_1\cdots X_n}$ if for $A \subset \mathbb{R}^n$,

$$\wp(X \in A) \;=\; \int_A f_X(x)\,dx \;=\; \int_{\mathbb{R}^n} I_A(x)\, f_X(x)\,dx,$$

which is shorthand for

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_A(x_1, \ldots, x_n)\, f_{X_1 \cdots X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Usually, we need to compute probabilities of the form

$$\wp(X_1 \in B_1, \ldots, X_n \in B_n) \;:=\; \wp\Big(\bigcap_{k=1}^{n} \{X_k \in B_k\} \Big),$$

where $B_1, \ldots, B_n$ are one-dimensional sets. The key to this calculation is to observe that

$$\bigcap_{k=1}^{n} \{X_k \in B_k\} \;=\; \Big\{ [X_1, \ldots, X_n]' \in B_1 \times \cdots \times B_n \Big\}.$$

(Recall that a vector $[x_1, \ldots, x_n]'$ lies in the Cartesian product set $B_1 \times \cdots \times B_n$ if and only if $x_i \in B_i$ for $i = 1, \ldots, n$.) It now follows that

$$\wp(X_1 \in B_1, \ldots, X_n \in B_n) \;=\; \int_{B_1} \cdots \int_{B_n} f_{X_1 \cdots X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

For example, to compute the joint cumulative distribution function, take $B_k = (\Leftrightarrow\infty, x_k]$ to get

$$\begin{aligned} F_X(x) &:= F_{X_1 \cdots X_n}(x_1, \ldots, x_n) \\ &:= \wp(X_1 \le x_1, \ldots, X_n \le x_n) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1 \cdots X_n}(t_1, \ldots, t_n)\, dt_1 \cdots dt_n, \end{aligned}$$

where we have changed the dummy variables of integration to t to avoid confusion with the upper limits of integration. Note also that

$$\left. \frac{\partial^n F_X}{\partial x_n \cdots \partial x_1} \right|_{x_1, \ldots, x_n} \;=\; f_{X_1 \cdots X_n}(x_1, \ldots, x_n).$$

We can use these equations to characterize the joint cdf and density of independent random variables. First, suppose $X_1, \ldots, X_n$ are independent. Then

$$\begin{aligned} F_X(x) &= \wp(X_1 \le x_1, \ldots, X_n \le x_n) \\ &= \wp(X_1 \le x_1) \cdots \wp(X_n \le x_n). \end{aligned}$$

By taking partial derivatives, we learn that

$$f_{X_1 \cdots X_n}(x_1, \ldots, x_n) \;=\; f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Thus, if the X_i are independent, then the joint density factors. On the other hand, if the joint density factors, then $\wp(X_1 \in B_1, \ldots, X_n \in B_n)$ has the form

$$\int_{B_1} \cdots \int_{B_n} f_{X_1}(x_1) \cdots f_{X_n}(x_n)\, dx_1 \cdots dx_n,$$

which factors into the product

$$\left(\int_{B_1} f_{X_1}(x_1)\,dx_1\right)\cdots\left(\int_{B_n} f_{X_n}(x_n)\,dx_n\right).$$

Thus, if the joint density factors,

$$\wp(X_1 \in B_1, \ldots, X_n \in B_n) = \wp(X_1 \in B_1)\cdots\wp(X_n \in B_n),$$

and we see that the X_i are independent.

Sometimes we need to compute probabilities involving only some of the X_i. In this case, we can obtain the joint density of the ones we need by integrating out the ones we do not need. For example,

$$f_{X_i}(x_i) \;=\; \int_{\mathbb{R}^{n-1}} f_X(x_1,\ldots,x_n)\,dx_1\cdots dx_{i-1}dx_{i+1}\cdots dx_n,$$

and

$$f_{X_1X_2}(x_1,x_2) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1\cdots X_n}(x_1,\ldots,x_n)\,dx_3\cdots dx_n.$$

We can use these results to compute conditional densities such as

$$f_{X_3\cdots X_n|X_1X_2}(x_3,\ldots,x_n|x_1,x_2) \;=\; \frac{f_{X_1\cdots X_n}(x_1,\ldots,x_n)}{f_{X_1X_2}(x_1,x_2)}.$$

***Example* 5.13.** Let

$$f_{XYZ}(x,y,z) \;=\; \frac{3z^2}{7\sqrt{2\pi}}e^{-zy}\exp\left[\text{-}\tfrac{1}{2}\left(\tfrac{x-y}{z}\right)^2\right],$$

for $y \geq 0$ and $1 \leq z \leq 2$, and $f_{XYZ}(x,y,z) = 0$ otherwise. Find $f_{YZ}(y,z)$ and $f_{X|YZ}(x|y,z)$. Then find $f_Z(z)$, $f_{Y|Z}(y|z)$, and $f_{XY|Z}(x,y|z)$.

Solution. Observe that the joint density can be written as

$$f_{XYZ}(x,y,z) \;=\; \frac{\exp\left[\text{-}\tfrac{1}{2}\left(\tfrac{x-y}{z}\right)^2\right]}{\sqrt{2\pi}\,z}\cdot ze^{-zy}\cdot\tfrac{3}{7}z^2.$$

The first factor as a function of x is an $N(y,z^2)$ density. Hence,

$$f_{YZ}(y,z) \;=\; \int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\,dx \;=\; ze^{-zy}\cdot\tfrac{3}{7}z^2,$$

and

$$f_{X|YZ}(x|y,z) \;=\; \frac{f_{XYZ}(x,y,z)}{f_{YZ}(y,z)} \;=\; \frac{\exp\left[\text{-}\tfrac{1}{2}\left(\tfrac{x-y}{z}\right)^2\right]}{\sqrt{2\pi}\,z}.$$

Thus, $f_{X|YZ}(\cdot|y,z) \sim N(y,z^2)$. Next, in the above formula for $f_{YZ}(y,z)$, observe that ze^{-zy} as a function of y is an exponential density with parameter z. Thus,

$$f_Z(z) \;=\; \int_0^\infty f_{YZ}(y,z)\,dy \;=\; \tfrac{3}{7}z^2, \quad 1 \le z \le 2.$$

It follows that $f_{Y|Z}(y|z) = f_{YZ}(y,z)/f_Z(z) = ze^{-zy}$; i.e., $f_{Y|Z}(\cdot|z) \sim \exp(z)$. Finally,

$$f_{XY|Z}(x,y|z) = \frac{f_{XYZ}(x,y,z)}{f_Z(z)} = \frac{\exp\left[\Leftrightarrow\frac{1}{2}\left(\frac{x-y}{z}\right)^2\right]}{\sqrt{2\pi}\,z} \cdot ze^{-zy}.$$

The preceding example shows that

$$f_{XYZ}(x,y,z) \;=\; f_{X|YZ}(x|y,z)\,f_{Y|Z}(y|z)\,f_Z(z).$$

More generally,

$$\begin{aligned} f_{X_1\cdots X_n}(x_1,\ldots,x_n) \;&=\; f_{X_1}(x_1)\,f_{X_2|X_1}(x_2|x_1)\,f_{X_3|X_1X_2}(x_3|x_1,x_2)\cdots \\ &\quad \cdots f_{X_n|X_1\cdots X_{n-1}}(x_n|x_1,\ldots,x_{n-1}). \end{aligned}$$

Contemplating this equation for a moment reveals that

$$f_{X_1\cdots X_n}(x_1,\ldots,x_n) = f_{X_1\cdots X_{n-1}}(x_1,\ldots,x_{n-1})\,f_{X_n|X_1\cdots X_{n-1}}(x_n|x_1,\ldots,x_{n-1}),$$

which is an important recursion that is discussed later.

If $g(x) = g(x_1,\ldots,x_n)$ is a real-valued function of the vector $x = [x_1,\ldots,x_n]'$, then we can compute

$$\mathsf{E}[g(X)] \;=\; \int_{\mathbb{R}^n} g(x)\,f_X(x)\,dx,$$

which is shorthand for

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n)\,f_{X_1\cdots X_n}(x_1,\ldots,x_n)\,dx_1\cdots dx_n.$$

The Law of Total Probability

We also have the law of total probability and the substitution law for multiple random variables. Let $U = [U_1,\ldots,U_m]'$ and $V = [V_1,\ldots,V_n]'$ be any two vectors of random variables. The law of total probability tells us that

$$\mathsf{E}[g(U,V)] \;=\; \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{m \text{ times}} \mathsf{E}[g(U,V)|U=u]\,f_U(u)\,du,$$

where, by the substitution law, $\mathsf{E}[g(U,V)|U=u] = \mathsf{E}[g(u,V)|U=u]$, and

$$\mathsf{E}[g(u,V)|U=u] \;=\; \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n \text{ times}} g(u,v)\, f_{V|U}(v|u)\, dv,$$

and

$$f_{V|U}(v|u) \;=\; \frac{f_{UV}(u_1,\ldots,u_m,v_1,\ldots,v_n)}{f_U(u_1,\ldots,u_m)}.$$

When g has a product form, say $g(u,v)=h(u)k(v)$, it is easy to see that

$$\mathsf{E}[h(u)k(V)|U=u] \;=\; h(u)\,\mathsf{E}[k(V)|U=u].$$

***Example* 5.14.** Let X, Y, and Z be as in Example 5.13. Find $\mathsf{E}[X]$ and $\mathsf{E}[XZ]$.

Solution. Rather than use the marginal density of X to compute $\mathsf{E}[X]$, we use the law of total probability. Write

$$\mathsf{E}[X] \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathsf{E}[X|Y=y,Z=z]\, f_{YZ}(y,z)\, dy\, dz.$$

Next, from Example 5.13, $f_{X|YZ}(\cdot|y,z) \sim N(y,z^2)$, and so $\mathsf{E}[X|Y=y,Z=z]=y$. Thus,

$$\begin{aligned}
\mathsf{E}[X] &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y\, f_{YZ}(y,z)\, dy\right) dz \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y\, f_{Y|Z}(y|z)\, dy\right) f_Z(z)\, dz \\
&= \int_{-\infty}^{\infty} \mathsf{E}[Y|Z=z]\, f_Z(z)\, dz.
\end{aligned}$$

From Example 5.13, $f_{Y|Z}(\cdot|z) \sim \exp(z)$ and $f_Z(z) = 3z^2/7$; hence, $\mathsf{E}[Y|Z=z]=1/z$, and we have

$$\mathsf{E}[X] \;=\; \int_1^2 \tfrac{3}{7} z\, dz \;=\; \tfrac{9}{14}.$$

Now, to find $\mathsf{E}[XZ]$, write

$$\mathsf{E}[XZ] \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathsf{E}[Xz|Y=y,Z=z]\, f_{YZ}(y,z)\, dy\, dz.$$

We then note that $\mathsf{E}[Xz|Y=y,Z=z] = \mathsf{E}[X|Y=y,Z=z]\, z = yz$. Thus,

$$\mathsf{E}[XZ] \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yz\, f_{YZ}(y,z)\, dy\, dz \;=\; \mathsf{E}[YZ].$$

In Problem 34 the reader is asked to show that $\mathsf{E}[YZ]=1$. Thus, $\mathsf{E}[XZ]=1$ as well.

***Example* 5.15.** Let N be a positive, integer-valued random variable, and let $X_1, X_2, \ldots$ be i.i.d. Further assume that N is independent of this sequence. Consider the **random sum**,

$$\sum_{i=1}^{N} X_i.$$

Note that the number of terms in the sum is a random variable. Find the mean value of the random sum.

Solution. Use the law of total probability to write

$$\mathsf{E}\Big[\sum_{i=1}^{N} X_i\Big] \;=\; \sum_{n=1}^{\infty} \mathsf{E}\Big[\sum_{i=1}^{n} X_i \Big| N = n\Big]\, \wp(N=n).$$

By independence of N and the X_i sequence,

$$\mathsf{E}\Big[\sum_{i=1}^{n} X_i \Big| N = n\Big] \;=\; \mathsf{E}\Big[\sum_{i=1}^{n} X_i\Big] \;=\; \sum_{i=1}^{n} \mathsf{E}[X_i].$$

Since the X_i are i.i.d., they all have the same mean. In particular, for all i, $\mathsf{E}[X_i] = \mathsf{E}[X_1]$. Thus,

$$\mathsf{E}\Big[\sum_{i=1}^{n} X_i \Big| N = n\Big] \;=\; n\,\mathsf{E}[X_1].$$

Now we can write

$$\begin{aligned} \mathsf{E}\Big[\sum_{i=1}^{N} X_i\Big] &= \sum_{n=1}^{\infty} n\,\mathsf{E}[X_1]\,\wp(N=n) \\ &= \mathsf{E}[N]\,\mathsf{E}[X_1]. \end{aligned}$$

5.6. Notes

Notes §5.1: Joint and Marginal Probabilities

***Note* 1.** Comments analogous to Note 1 in Chapter 2 apply here. Specifically, the set A must be restricted to a suitable σ-field $\mathcal{B}$ of subsets of $\mathbb{R}^2$. Typically, $\mathcal{B}$ is taken to be the collection of Borel sets of $\mathbb{R}^2$; i.e., $\mathcal{B}$ is the smallest σ-field containing all the open sets of $\mathbb{R}^2$.

***Note* 2.** We now derive the limit formula for $F_X(x)$ in (5.1); the formula for $F_Y(y)$ can be derived similarly. To begin, write

$$F_X(x) \;:=\; \wp(X \le x) \;=\; \wp((X,Y) \in (\Leftrightarrow\infty, x] \times \mathbb{R}).$$

Next, observe that $\mathbb{R} = \bigcup_{n=1}^{\infty}(\Leftrightarrow\infty, n]$, and write

$$\begin{aligned}(\Leftrightarrow\infty, x] \times \mathbb{R} &= (\Leftrightarrow\infty, x] \times \bigcup_{n=1}^{\infty}(\Leftrightarrow\infty, n] \\ &= \bigcup_{n=1}^{\infty}(\Leftrightarrow\infty, x] \times (\Leftrightarrow\infty, n].\end{aligned}$$

Since the union is increasing, we can use the limit property (1.4) to show that

$$\begin{aligned}F_X(x) &= \wp\left((X,Y) \in \bigcup_{n=1}^{\infty}(\Leftrightarrow\infty, x] \times (\Leftrightarrow\infty, n]\right) \\ &= \lim_{N\to\infty} \wp((X,Y) \in (\Leftrightarrow\infty, x] \times (\Leftrightarrow\infty, N]) \\ &= \lim_{N\to\infty} F_{XY}(x, N).\end{aligned}$$

Notes §5.3: Conditional Probability and Expectation

***Note* 3.** To show that the law of substitution holds for conditional probability, write

$$\wp(g(X,Y) \in C) = \mathsf{E}[I_C(g(X,Y))] = \int_{-\infty}^{\infty} \mathsf{E}[I_C(g(X,Y))|X = x] f_X(x)\, dx$$

and reduce the problem to one involving conditional expectation, for which the law of substitution has already been established.

5.7. Problems

Problems §5.1: Joint and Marginal Distributions

1. For $a < b$ and $c < d$, sketch the following sets.
   (a) $R := (a, b] \times (c, d]$.
   (b) $A := (\Leftrightarrow\infty, a] \times (\Leftrightarrow\infty, d]$.
   (c) $B := (\Leftrightarrow\infty, b] \times (\Leftrightarrow\infty, c]$.
   (d) $C := (a, b] \times (\Leftrightarrow\infty, c]$.
   (e) $D := (\Leftrightarrow\infty, a] \times (c, d]$.
   (f) $A \cap B$.
2. Show that $\wp(a < X \le b, c < Y \le d)$ is given by
   $$F_{XY}(b,d) \Leftrightarrow F_{XY}(a,d) \Leftrightarrow F_{XY}(b,c) + F_{XY}(a,c).$$
   Hint: Using the notation of the preceding problem, observe that
   $$(\Leftrightarrow\infty, b] \times (\Leftrightarrow\infty, d] = R \cup (A \cup B),$$
   and solve for $\wp((X,Y) \in R)$.

3. The joint density in Example 5.4 was obtained by differentiating $F_{XY}(x,y)$ first with respect to x and then with respect to y. In this problem, find the joint density by differentiating first with respect to y and then with respect to x.

4. Find the marginals $F_X(x)$ and $F_Y(y)$ if

$$F_{XY}(x,y) = \begin{cases} x \Leftrightarrow 1 \Leftrightarrow \dfrac{e^{-y} \Leftrightarrow e^{-xy}}{y}, & 1 \le x \le 2,\ y > 0, \\ 1 \Leftrightarrow \dfrac{e^{-y} \Leftrightarrow e^{-2y}}{y}, & x > 2,\ y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

5. Find the marginals $F_X(x)$ and $F_Y(y)$ if

$$F_{XY}(x,y) = \begin{cases} \frac{2}{7}(1 \Leftrightarrow e^{-2y}), & 2 \le x < 3,\ y \ge 0, \\ \dfrac{(7 \Leftrightarrow 2e^{-2y} \Leftrightarrow 5e^{-3y})}{7}, & x \ge 3,\ y \ge 0, \\ 0, & \text{otherwise.} \end{cases}$$

Problems §5.2: Jointly Continuous Random Variables

6. Find the marginal density $f_X(x)$ if

$$f_{XY}(x,y) = \frac{\exp[\Leftrightarrow|y \Leftrightarrow x| \Leftrightarrow x^2/2]}{2\sqrt{2\pi}}.$$

7. Find the marginal density $f_Y(y)$ if

$$f_{XY}(x,y) = \frac{4e^{-(x-y)^2/2}}{y^5\sqrt{2\pi}}, \quad y \ge 1.$$

8. Let X and Y have joint density $f_{XY}(x,y)$. Find the marginal cdf and density of $\max(X,Y)$ and of $\min(X,Y)$. How do your results simplify if X and Y are independent? What if you further assume that the densities of X and Y are the same?

9. Let X and Y be independent gamma random variables with positive parameters p and q, respectively. Find the density of $Z := X + Y$. Then compute $\wp(Z > 1)$ if $p = q = 1/2$.

10. Find the density of $Z := X + Y$, where X and Y are independent Cauchy random variables with parameters λ and μ, respectively. Then compute $\wp(Z \le 1)$ if $\lambda = \mu = 1/2$.

*11. If $X \sim N(0,1)$, then the complementary cumulative distribution function (ccdf) of X is

$$Q(x_0) \;:=\; \wp(X > x_0) \;=\; \int_{x_0}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx.$$

(a) Show that

$$Q(x_0) \;=\; \frac{1}{\pi}\int_0^{\pi/2} \exp\left(\frac{-x_0^2}{2\cos^2\theta}\right) d\theta, \quad x_0 \geq 0.$$

Hint: For any random variables X and Y, we can always write

$$\wp(X > x_0) = \wp(X > x_0, Y \in \mathbb{R}) = \wp((X,Y) \in D),$$

where D is the half plane $D := \{(x,y) : x > x_0\}$. Now specialize to the case where X and Y are independent and both $N(0,1)$. Then the probability on the right is a double integral that can be evaluated using polar coordinates.

Remark. The procedure outlined in the hint is a generalization of that used in Section 3.1 to show that the standard normal density integrates to one. To see this, note that if $x_0 = -\infty$, then $D = \mathbb{R}^2$.

(b) Use the result of (a) to derive **Craig's formula** [10, p. 572, eq. (9)],

$$Q(x_0) \;=\; \frac{1}{\pi}\int_0^{\pi/2} \exp\left(\frac{-x_0^2}{2\sin^2 t}\right) dt, \quad x_0 \geq 0.$$

Remark. Simon [41] has derived a similar result for the Marcum Q function (defined in Problem 17 in Chapter 4) and its higher-order generalizations. See also [43, pp. 1865–1867].

Problems §5.3: Conditional Probability and Expectation

12. Let $f_{XY}(x,y)$ be as derived in Example 5.4, and note that $f_X(x)$ and $f_Y(y)$ were found in Example 5.5. Find $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ for $x, y > 0$.

13. Let $f_{XY}(x,y)$ be as derived in Example 5.4, and note that $f_X(x)$ and $f_Y(y)$ were found in Example 5.5. Compute $\mathsf{E}[Y|X = x]$ for $x > 0$ and $\mathsf{E}[X|Y = y]$ for $y > 0$.

14. Let X and Y be jointly continuous. Show that if

$$\wp(Y \in C|X = x) \;:=\; \int_C f_{Y|X}(y|x)\, dy,$$

then

$$\wp(Y \in C) \;=\; \int_{-\infty}^{\infty} \wp(Y \in C|X = x) f_X(x)\, dx.$$

15. Find $\wp(X \le Y)$ if X and Y are independent with $X \sim \exp(\lambda)$ and $Y \sim \exp(\mu)$.

16. Let X and Y be independent random variables with Y being exponential with parameter 1 and X being uniform on $[1,2]$. Find $\wp(Y/\ln(1+X^2) > 1)$.

17. Let X and Y be jointly continuous random variables with joint density f_{XY}. Find $f_Z(z)$ if

    (a) $Z = e^X Y$.

    (b) $Z = |X+Y|$.

18. Let X and Y be independent continuous random variables with respective densities f_X and f_Y. Put $Z = Y/X$.

    (a) Find the density of Z. *Hint:* Review Example 5.9.

    (b) If X and Y are both $N(0,\sigma^2)$, show that Z has a Cauchy(1) density that does not depend on σ^2.

    (c) If X and Y are both Laplace(λ), find a closed-form expression for $f_Z(z)$ that does not depend on λ.

    (d) Find a closed-form expression for the density of Z if Y is uniform on $[-1,1]$ and $X \sim N(0,1)$.

    (e) If X and Y are both Rayleigh random variables with parameter λ, find a closed-form expression for the density of Z. Your answer should not depend on λ.

19. Let X and Y be independent with densities $f_X(x)$ and $f_Y(y)$. If X is a positive random variable, and if $Z = Y/\ln(X)$, find the density of Z.

20. Let $Y \sim \exp(\lambda)$, and suppose that given $Y = y$, $X \sim \text{gamma}(p,y)$. Assuming $r > n$, evaluate $\mathsf{E}[X^n Y^r]$.

21. Use the law of total probability to to solve the following problems.

    (a) Evaluate $\mathsf{E}[\cos(X+Y)]$ if given $X = x$, Y is conditionally uniform on $[x-\pi, x+\pi]$.

    (b) Evaluate $\wp(Y > y)$ if $X \sim \text{uniform}[1,2]$, and given $X = x$, Y is exponential with parameter x.

    (c) Evaluate $\mathsf{E}[Xe^Y]$ if $X \sim \text{uniform}[3,7]$, and given $X = x$, $Y \sim N(0,x^2)$.

    (d) Let $X \sim \text{uniform}[1,2]$, and suppose that given $X = x$, $Y \sim N(0,1/x)$. Evaluate $\mathsf{E}[\cos(XY)]$.

22. Find $\mathsf{E}[X^n Y^m]$ if $Y \sim \exp(\beta)$, and given $Y = y$, $X \sim \text{Rayleigh}(y)$.

*23. Let $X \sim \text{gamma}(p,\lambda)$ and $Y \sim \text{gamma}(q,\lambda)$ be independent.

(a) If $Z := X/Y$, show that the density of Z is

$$f_Z(z) = \frac{1}{B(p,q)} \cdot \frac{z^{p-1}}{(1+z)^{p+q}}, \quad z > 0.$$

Observe that $f_Z(z)$ depends on p and q, but not on λ. It was shown in Problem 19 in Chapter 3 that $f_Z(z)$ integrates to one. *Hint:* You will need the fact that $B(p,q) = , (p), (q)/, (p+q)$, which was shown in Problem 13 in Chapter 3.

(b) Show that $V := X/(X+Y)$ has a beta density with parameters p and q. *Hint:* Observe that $V = Z/(1+Z)$, where $Z = X/Y$ as above.

Remark. If $W := (X/p)/(Y/q)$, then $f_W(w) = (p/q)\, f_Z(w(p/q))$. If further $p = k_1/2$ and $q = k_2/2$, then W is said to be an ***F* random variable** with k_1 and k_2 degrees of freedom. If further $\lambda = 1/2$, then the X and Y are chi-squared with k_1 and k_2 degrees of freedom, respectively.

*24. Let X and Y be independent with $X \sim N(0,1)$ and Y being chi-squared with k degrees of freedom. Show that the density of $Z := X/\sqrt{Y/k}$ has a student's t density with k degrees of freedom. *Hint:* For this problem, it may be helpful to review the results of Problems 11–13 and 17 in Chapter 3.

Problems §5.4: The Bivariate Normal

25. Let U and V have the joint Gaussian density in (5.11). Show that for all ρ with $\Leftrightarrow 1 < \rho < 1$, U and V both have standard univariate $N(0,1)$ marginal densities that do not involve ρ.

26. Let X and Y be jointly Gaussian with density $f_{XY}(x,y)$ given by (5.13). Find $f_X(x)$, $f_Y(y)$, $f_{X|Y}(x|y)$, and $f_{Y|X}(y|x)$.

27. Let X and Y be jointly Gaussian with density $f_{XY}(x,y)$ given by (5.13). Find $\mathsf{E}[Y|X=x]$ and $\mathsf{E}[X|Y=y]$.

28. If X and Y are jointly normal with parameters, m_X, m_Y, σ_X^2, σ_Y^2, and ρ. Compute $\mathsf{E}[X]$, $\mathsf{E}[X^2]$, and $\mathsf{E}[XY]$.

*29. Let φ_ρ be the standard bivariate normal density defined in (5.11). Put

$$f_{UV}(u,v) := \frac{1}{2}[\varphi_{\rho_1}(u,v) + \varphi_{\rho_2}(u,v)],$$

where $\Leftrightarrow 1 < \rho_1 \neq \rho_2 < 1$.

(a) Show that the marginals f_U and f_V are both $N(0,1)$. (You may use the results of Problem 25.)

(b) Show that $\overline{\rho} := \mathsf{E}[UV] = (\rho_1 + \rho_2)/2$. (You may use the result of Example 5.10.)

(c) Show that U and V cannot be jointly normal. *Hints:* (*i*) To obtain a contradiction, suppose that f_{UV} is a jointly normal density with parameters given by parts (a) and (b). (*ii*) Consider $f_{UV}(u,u)$. (*iii*) Use the following fact: If $\beta_1, \ldots, \beta_n$ are distinct real numbers, and if

$$\sum_{k=1}^{n} \alpha_k e^{\beta_k t} \;=\; 0, \quad \text{for all } t \geq 0,$$

then $\alpha_1 = \cdots = \alpha_n = 0$.

⋆30. Let U and V be jointly normal with joint density $\varphi_\rho(u,v)$ defined in (5.11). Put

$$Q_\rho(u_0, v_0) \;:=\; \wp(U > u_0, V > v_0).$$

Show that for $u_0, v_0 \geq 0$,

$$\begin{aligned} Q_\rho(u_0, v_0) \;&=\; \int_0^{\pi/2 - \tan^{-1}(u_0/v_0)} h_\rho(u_0^2, \theta)\, d\theta \\ &\quad + \int_0^{\tan^{-1}(u_0/v_0)} h_\rho(v_0^2, \theta)\, d\theta, \end{aligned}$$

where

$$h_\rho(z,\theta) \;:=\; \frac{\sqrt{1-\rho^2}}{2\pi(1-\rho\sin 2\theta)} \exp\left[\frac{-z(1-\rho\sin 2\theta)}{2(1-\rho^2)\sin^2\theta}\right].$$

This formula for $Q_\rho(u_0, v_0)$ is Simon's bivariate generalization [43, pp. 1864–1865] of Craig's univariate formula given in Problem 11. *Hint:* Write $\wp(U > u_0, V > v_0)$ as a double integral and convert to polar coordinates. It may be helpful to review your solution of Problem 11 first.

⋆31. Use Simon's formula (Problem 30) to show that

$$Q(x_0)^2 \;=\; \frac{1}{\pi} \int_0^{\pi/4} \exp\left(\frac{-x_0^2}{2\sin^2 t}\right) dt.$$

In other words, to compute $Q(x_0)^2$, we integrate Craig's integrand (Problem 11) only half as far [42, p. 210]!

Problems §5.5: ⋆Multivariate Random Variables

⋆32. If

$$f_{XYZ}(x,y,z) = \frac{2\exp[-|x-y| - (y-z)^2/2]}{z^5\sqrt{2\pi}}, \quad z \geq 1,$$

and $f_{XYZ}(x,y,z) = 0$ otherwise, find $f_{YZ}(y,z)$, $f_{X|YZ}(x|y,z)$, $f_Z(z)$, and $f_{Y|Z}(y|z)$.

*33. Let

$$f_{XYZ}(x,y,z) \;=\; \frac{e^{-(x-y)^2/2}e^{-(y-z)^2/2}e^{-z^2/2}}{(2\pi)^{3/2}}.$$

Find $f_{XY}(x,y)$. Then find the means and variances of X and Y. Also find the correlation, $\mathsf{E}[XY]$.

*34. Let X, Y, and Z be as in Example 5.13. Evaluate $\mathsf{E}[XY]$ and $\mathsf{E}[YZ]$.

*35. Let X, Y, and Z be as in Problem 32. Evaluate $\mathsf{E}[XYZ]$.

*36. Let X, Y, and Z be jointly continuous. Assume that $X \sim \text{uniform}[1,2]$; that given $X = x$, $Y \sim \exp(1/x)$; and that given $X = x$ and $Y = y$, Z is $N(x,1)$. Find $\mathsf{E}[XYZ]$.

*37. Let N denote the number of primaries in a photomultiplier, and let X_i be the number of secondaries due to the ith primary. Then the total number of secondaries is

$$Y \;=\; \sum_{i=1}^{N} X_i.$$

Express the characteristic function of Y in terms of the probability generating function of N, $G_N(z)$, and the characteristic function of the X_i, assuming that the X_i are i.i.d. with common characteristic function $\varphi_X(\nu)$. Assume that N is independent of the X_i sequence. Find the density of Y if $N \sim \text{geometric}_1(p)$ and $X_i \sim \exp(\lambda)$.

CHAPTER 6
Introduction to Random Processes

A **random process** or **stochastic process** is any family of random variables. For example, consider the experiment consisting of three coin tosses. A suitable sample space would be

$$\Omega := \{\text{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH}\}.$$

On this sample space, we can define several random variables. For $n = 1, 2, 3$, put

$$X_n(\omega) := \begin{cases} 0, & \text{if the } n\text{th component of } \omega \text{ is T,} \\ 1, & \text{if the } n\text{th component of } \omega \text{ is H.} \end{cases}$$

Then, for example, $X_1(\text{THH}) = 0$, $X_2(\text{THH}) = 1$, and $X_3(\text{THH}) = 1$. For fixed ω, we can plot the sequence $X_1(\omega)$, $X_2(\omega)$, $X_3(\omega)$, called the **sample path** corresponding to ω. This is done for $\omega = \text{THH}$ and for $\omega = \text{HTH}$ in Figure 6.1. These two graphs illustrate the important fact that every time the sample point ω changes, the sample path also changes.

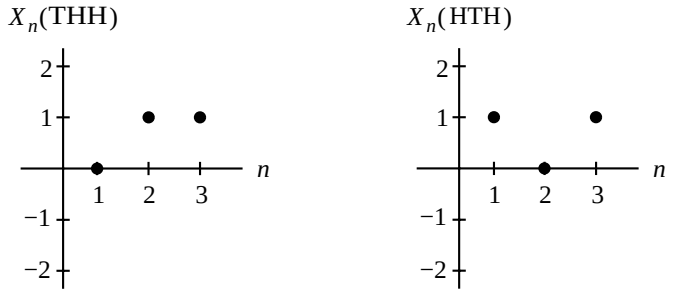

Figure 6.1. Sample paths $X_n(\omega)$ for $\omega = \text{THH}$ at the left and for $\omega = \text{HTH}$ at the right.

In general, a random process X_n may be defined over any range of integers n, which we usually think of as discrete time. In this case, we can usually plot only a portion of a sample path as in Figure 6.2. Notice that there is no requirement that X_n be integer valued. For example, in Figure 6.2, $X_0(\omega) = 0.5$ and $X_1(\omega) = 2.5$.

It is also useful to allow continuous-time random processes X_t where t can be any real number, not just an integer. Thus, for each t, $X_t(\omega)$ is a random variable, or function of ω. However, as in the discrete-time case, we can fix ω and allow the time parameter t to vary. In this case, for fixed ω, if we plot the sample path $X_t(\omega)$ as a function of t, we get a curve instead of a sequence as shown in Figure 6.3.

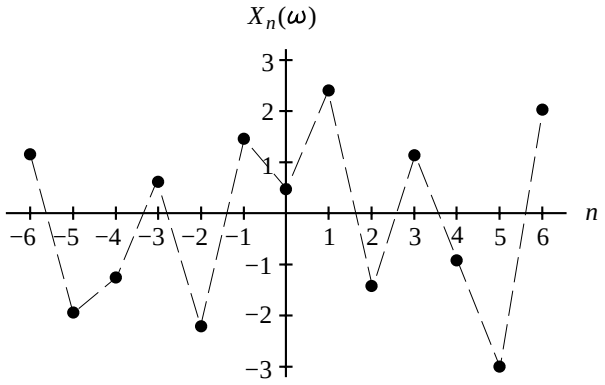

Figure 6.2. A portion of a sample path $X_n(\omega)$ of a discrete-time random process. Because there are more dots than in the previous figure, the dashed line has been added to more clearly show the ordering of the dots.

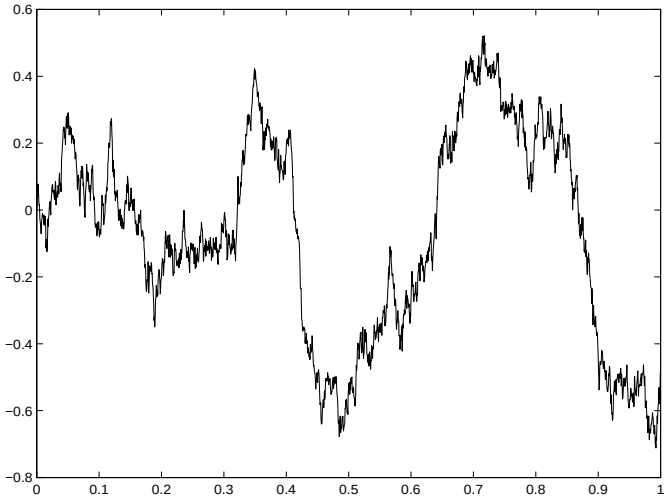

Figure 6.3. Sample path $X_t(\omega)$ of a continuous-time random process with continuous but very wiggly sample paths.

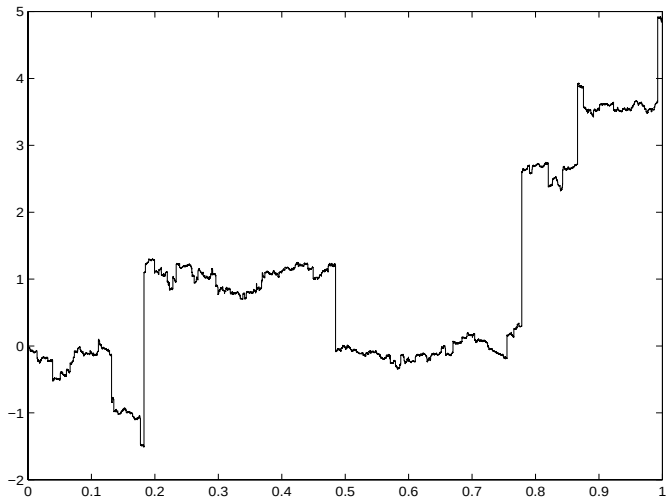

Figure 6.4. Sample path $X_t(\omega)$ of a continuous-time random process with very wiggly sample paths and jump discontinuities.

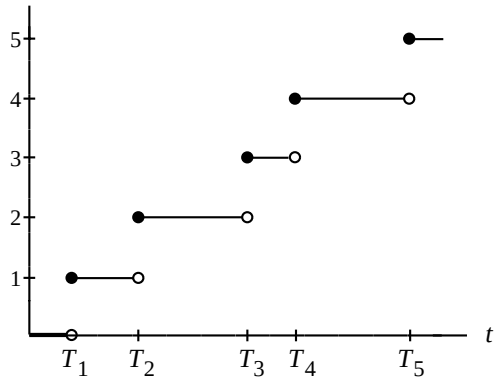

Figure 6.5. Sample path $X_t(\omega)$ of a continuous-time random process with sample paths that are piecewise constant with jumps at random times.

Figure 6.3 shows a very wiggly but continuous waveform. However, wiggly waveforms with jump discontinuities are also possible as shown in Figure 6.4.

Figures 6.3 and 6.4 are what we usually think of when we think of a random process as a model for noise waveforms. However, random waveforms do not have to be wiggly. They can be very regular as in Figure 6.5.

The main point of the foregoing pictures and discussion is to emphasize that there are two ways to think about a random process. The first way is to think of X_n or X_t as a family of random variables, which by definition, is just a family of functions of ω. The second way is to think of X_n or X_t as a random waveform in discrete or continuous time, respectively.

Section 6.1 introduces the mean function, the correlation function, and the covariance function of a random process. Section 6.2 introduces the concept of a stationary process and of a wide-sense stationary process. The correlation function and the power spectral density of wide-sense stationary processes are introduced and their properties are derived. Section 6.3 is the heart of the chapter, and covers the analysis of wide-sense stationary processes through linear, time-invariant systems. These results are then applied in Sections 6.4 and 6.5 to derive the matched filter and the Wiener filter. Section 6.6 contains a discussion of the expected time-average power in a wide-sense stationary random process. The Wiener–Khinchin Theorem is derived and used to provide an alternative expression for the power spectral density. As an easy corollary of our derivation of the Wiener–Khinchin Theorem, we obtain the mean-square ergodic theorem for wide-sense stationary processes. Section 6.7 briefly extends the notion of power spectral density to processes that are not wide-sense stationary.

6.1. Mean, Correlation, and Covariance

If X_t is a random process, its **mean function** is

$$m_X(t) := \mathsf{E}[X_t].$$

Its **correlation function** is

$$R_X(t_1,t_2) := \mathsf{E}[X_{t_1}X_{t_2}].$$

***Example* 6.1.** In modeling a communication system, the carrier signal at the receiver is modeled by $X_t = \cos(2\pi ft + \Theta)$, where $\Theta \sim \text{uniform}[-\pi,\pi]$. The reason for the random phase is that the receiver does not know the exact time when the transmitter was turned on. Find the mean function and the correlation function of X_t.

Solution. For the mean, write

$$\begin{aligned}\mathsf{E}[X_t] &= \mathsf{E}[\cos(2\pi ft+\Theta)]\\ &= \int_{-\infty}^{\infty}\cos(2\pi ft+\theta)\,f_\Theta(\theta)\,d\theta\\ &= \int_{-\pi}^{\pi}\cos(2\pi ft+\theta)\frac{d\theta}{2\pi}.\end{aligned}$$

Be careful to observe that this last integral is with respect to θ, *not* t. Hence, this integral evaluates to zero.

For the correlation, write

$$\begin{aligned} R_X(t_1,t_2) &= \mathsf{E}[X_{t_1}X_{t_2}] \\ &= \mathsf{E}\big[\cos(2\pi f t_1+\Theta)\cos(2\pi f t_2+\Theta)\big] \\ &= \frac{1}{2}\mathsf{E}\big[\cos(2\pi f[t_1+t_2]+2\Theta)+\cos(2\pi f[t_1-t_2])\big]. \end{aligned}$$

The first cosine has expected value zero just as the mean did. The second cosine is nonrandom, and equal to its expected value. Thus, $R_X(t_1,t_2) = \cos(2\pi f[t_1-t_2])/2$.

Correlation functions have special properties. First,

$$R_X(t_1,t_2) \;=\; \mathsf{E}[X_{t_1}X_{t_2}] \;=\; \mathsf{E}[X_{t_2}X_{t_1}] \;=\; R_X(t_2,t_1).$$

In other words, the correlation function is a **symmetric function** of t_1 and t_2. Next, observe that $R_X(t,t) = \mathsf{E}[X_t^2] \geq 0$, and for any t_1 and t_2,

$$\big|R_X(t_1,t_2)\big| \;\leq\; \sqrt{\mathsf{E}[X_{t_1}^2]\,\mathsf{E}[X_{t_2}^2]}\,. \tag{6.1}$$

This is an easy consequence of the **Cauchy–Schwarz inequality** (see Problem 1), which says that for any random variables U and V,

$$\mathsf{E}[UV]^2 \;\leq\; \mathsf{E}[U^2]\,\mathsf{E}[V^2].$$

Taking $U = X_{t_1}$ and $V = X_{t_2}$ yields the desired result.

The **covariance function** is

$$C_X(t_1,t_2) \;:=\; \mathsf{E}\big[\big(X_{t_1}-\mathsf{E}[X_{t_1}]\big)\big(X_{t_2}-\mathsf{E}[X_{t_2}]\big)\big],$$

An easy calculation shows that

$$C_X(t_1,t_2) \;=\; R_X(t_1,t_2)-m_X(t_1)m_X(t_2).$$

Note that the covariance function is also symmetric; i.e., $C_X(t_1,t_2) = C_X(t_2,t_1)$.

Since we usually assume that our processes are zero mean; i.e., $m_X(t) \equiv 0$, we focus on the correlation function and its properties.

6.2. Wide-Sense Stationary Processes

Strict-Sense Stationarity

A random process is **strictly stationary** if for any finite collection of times $t_1,\ldots,t_n$, all joint probabilities involving $X_{t_1},\ldots,X_{t_n}$ are the same as those involving $X_{t_1+\Delta t},\ldots,X_{t_n+\Delta t}$ for any positive or negative time shift Δt. For discrete-time processes, this is equivalent to requiring that joint probabilities involving $X_1,\ldots,X_n$ are the same as those involving $X_{1+m},\ldots,X_{n+m}$ for any integer m.

***Example* 6.2.** Consider a discrete-time process of integer-valued random variables X_n such that

$$\wp(X_1 = i_1, \ldots, X_n = i_n) \;=\; q(i_1)\, r(i_2|i_1)\, r(i_3|i_2) \cdots r(i_n|i_{n-1}),$$

where $q(i)$ is any pmf, and for each i, $r(j|i)$ is a pmf in the variable j; i.e., r is any conditional pmf. Show that if q has the property*

$$\sum_k q(k) r(j|k) \;=\; q(j), \tag{6.2}$$

then X_n is strictly stationary for positive time shifts m.

Solution. Observe that

$$\wp(X_1 = j_1, \ldots, X_m = j_m, X_{1+m} = i_1, \ldots, X_{n+m} = i_n)$$

is equal to

$$q(j_1) r(j_2|j_1) r(j_3|j_2) \cdots r(i_1|j_m) r(i_2|i_1) \cdots r(i_n|i_{n-1}).$$

Summing both expressions over j_1 and using (6.2) shows that

$$\wp(X_2 = j_2, \ldots, X_m = j_m, X_{1+m} = i_1, \ldots, X_{n+m} = i_n)$$

is equal to

$$q(j_2) r(j_3|j_2) \cdots r(i_1|j_m) r(i_2|i_1) \cdots r(i_n|i_{n-1}).$$

Continuing in this way, summing over $j_2, \ldots, j_m$, shows that

$$\wp(X_{1+m} = i_1, \ldots, X_{n+m} = i_n) \;=\; q(i_1) r(i_2|i_1) \cdots r(i_n|i_{n-1}),$$

which was the definition of $\wp(X_1 = i_1, \ldots, X_n = i_n)$.

Strict stationarity is a very strong property with many implications. Even taking $n = 1$ in the definition tells us that for any t_1 and $t_1 + \Delta t$, X_{t_1} and $X_{t_1+\Delta t}$ have the same pmf or density. It then follows that for any function $g(x)$, $\mathsf{E}[g(X_{t_1})] = \mathsf{E}[g(X_{t_1+\Delta t})]$. Taking $\Delta t = \text{-}t_1$ shows that $\mathsf{E}[g(X_{t_1})] = \mathsf{E}[g(X_0)]$, which does not depend on t_1. Stationarity also implies that for any function $g(x_1, x_2)$, we have

$$\mathsf{E}[g(X_{t_1}, X_{t_2})] \;=\; \mathsf{E}[g(X_{t_1+\Delta t}, X_{t_2+\Delta t})]$$

for every time shift Δt. Since Δt is arbitrary, let $\Delta t = \text{-}t_2$. Then

$$\mathsf{E}[g(X_{t_1}, X_{t_2})] \;=\; \mathsf{E}[g(X_{t_1-t_2}, X_0)].$$

*In Chapter 9, we will see that X_n is a time-homogeneous Markov chain. The condition on q is that it be the equilibrium distribution of the chain.

It follows that $\mathsf{E}[g(X_{t_1}, X_{t_2})]$ depends on t_1 and t_2 only through the time difference $t_1 \Leftrightarrow t_2$.

Asking that all the joint probabilities involving any $X_{t_1}, \ldots, X_{t_n}$ be the same as those involving $X_{t_1+\Delta t}, \ldots, X_{t_n+\Delta t}$ for every time shift Δt is a very strong requirement. In practice, e.g., analyzing receiver noise in a communication system, it is often enough to require only that $\mathsf{E}[X_t]$ not depend on t and that the correlation $R_X(t_1, t_2) = \mathsf{E}[X_{t_1} X_{t_2}]$ depend on t_1 and t_2 only through the time difference, $t_1 \Leftrightarrow t_2$. This is a much weaker requirement than stationarity for several reasons. First, we are only concerned with one or two variables at a time rather than arbitrary finite collections. Second, we are not concerned with probabilities, only expectations. Third, we are only concerned with $\mathsf{E}[X_t]$ and $\mathsf{E}[X_{t_1} X_{t_2}]$ rather than $\mathsf{E}[g(X_t)]$ and $\mathsf{E}[g(X_{t_1}, X_{t_2})]$ for arbitrary functions g.

Wide-Sense Stationarity

We say that a process is **wide-sense stationary** (**WSS**) if the following two properties *both* hold:

(*i*) The mean function $m_X(t) = \mathsf{E}[X_t]$ does not depend on t.
(*ii*) The correlation function $R_X(t_1, t_2) = \mathsf{E}[X_{t_1} X_{t_2}]$ depends on t_1 and t_2 only through the time difference $t_1 \Leftrightarrow t_2$.

The process of Example 6.1 was WSS since we showed that $\mathsf{E}[X_t] = 0$ and $R_X(t_1, t_2) = \cos(2\pi f[t_1 \Leftrightarrow t_2])/2$.

Once we know a process is WSS, we write $R_X(t_1 \Leftrightarrow t_2)$ instead of $R_X(t_1, t_2)$. We can also write

$$R_X(\tau) \;=\; \mathsf{E}[X_{t+\tau} X_t], \quad \text{for any choice of } t.$$

In particular, note that

$$R_X(0) \;=\; \mathsf{E}[X_t^2] \;\geq\; 0, \quad \text{for all } t.$$

At time t, the instantaneous power in a WSS process is X_t^2.[†] The **expected instantaneous power** is then

$$P_X \;:=\; \mathsf{E}[X_t^2] \;=\; R_X(0),$$

which does not depend on t, since the process is WSS.

It is now convenient to introduce the Fourier transform of $R_X(\tau)$,

$$S_X(f) \;:=\; \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau}\, d\tau.$$

We call $S_X(f)$ the **power spectral density** of the process. Observe that by the Fourier inversion formula,

$$R_X(\tau) \;=\; \int_{-\infty}^{\infty} S_X(f) e^{j2\pi f\tau}\, df.$$

[†] If X_t is the voltage across a one-ohm resistor or the current passing through a one-ohm resistor, then X_t^2 is the instantaneous power dissipated.

In particular, taking $\tau = 0$, shows that

$$\int_{-\infty}^{\infty} S_X(f)\,df \;=\; R_X(0) \;=\; P_X.$$

To explain the terminology, "power spectral density," recall that *probability densities* are nonnegative functions that are integrated to obtain probabilities. Similarly, *power* spectral *densities* are nonnegative functions[‡] that are integrated to obtain powers. The adjective "spectral" refers to the fact that $S_X(f)$ is a function of frequency.

***Example* 6.3.** Let X_t be a WSS random process with power spectral density $S_X(f) = e^{-f^2/2}$. Find the power in the process.

Solution. The power is

$$\begin{aligned} P_X &= \int_{-\infty}^{\infty} S_X(f)\,df \\ &= \int_{-\infty}^{\infty} e^{-f^2/2}\,df \\ &= \sqrt{2\pi}\int_{-\infty}^{\infty} \frac{e^{-f^2/2}}{\sqrt{2\pi}}\,df \\ &= \sqrt{2\pi}. \end{aligned}$$

***Example* 6.4.** Let X_t be a WSS process with correlation function $R_X(\tau) = 5\sin(\tau)/\tau$. Find the power in the process.

Solution. The power is

$$P_X \;=\; R_X(0) \;=\; \lim_{\tau\to 0} R_X(\tau) \;=\; \lim_{\tau\to 0} 5\frac{\sin(\tau)}{\tau} \;=\; 5.$$

***Example* 6.5.** Let X_t be a WSS random process with ideal lowpass power spectral density $S_X(f) = I_{[-W,W]}(f)$ shown in Figure 6.6. Find its correlation function $R_X(\tau)$.

Solution. We must find the inverse Fourier transform of $S_X(f)$. Write

$$\begin{aligned} R_X(\tau) &= \int_{-\infty}^{\infty} S_X(f)e^{j2\pi f\tau}\,df \\ &= \int_{-W}^{W} e^{j2\pi f\tau}\,df \end{aligned}$$

[‡]The nonnegativity of power spectral densities is derived in Example 6.9.

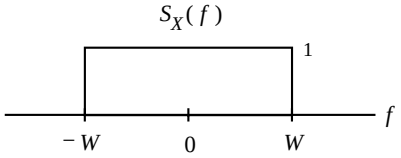

Figure 6.6. Power spectral density of bandlimited noise in Example 6.5.

$$
\begin{aligned}
&= \left.\frac{e^{j2\pi f\tau}}{j2\pi\tau}\right|_{f=-W}^{f=W} \\
&= \frac{e^{j2\pi W\tau} \Leftrightarrow e^{-j2\pi W\tau}}{j2\pi\tau} \\
&= 2W\frac{e^{j2\pi W\tau} \Leftrightarrow e^{-j2\pi W\tau}}{2j(2\pi W\tau)} \\
&= 2W\frac{\sin(2\pi W\tau)}{2\pi W\tau}.
\end{aligned}
$$

If the bandwidth W of $S_X(f)$ in the preceding example is allowed to go to infinity, then all frequencies will be present in equal amounts. A WSS process whose power spectral density is constant for all f is call **white noise**. The value of the constant is usually denoted by $\mathcal{N}_0/2$. Since the inverse transform of a constant function is a delta function, the correlation function of white noise with $S_X(f) = \mathcal{N}_0/2$ is $R_X(\tau) = \frac{\mathcal{N}_0}{2}\delta(\tau)$.

Remark. Just as the delta function is not an ordinary function, white noise is not an ordinary random process. For example, since $\delta(0)$ is not defined, and since $\mathsf{E}[X_t^2] = R_X(0) = \frac{\mathcal{N}_0}{2}\delta(0)$, we cannot speak of the second moment of X_t when X_t is white noise. On the other hand, since $S_X(f) = \mathcal{N}_0/2$ for white noise, and since

$$\int_{-\infty}^{\infty} \frac{\mathcal{N}_0}{2}\, df \;=\; \infty,$$

we often say that white noise has infinite average power.

Properties of Correlation Functions and Power Spectral Densities

The correlation function $R_X(\tau)$ is completely characterized by the following three properties.[1]

(*i*) $R_X(\Leftrightarrow\tau) = R_X(\tau)$; i.e., R_X is an even function of τ.
(*ii*) $|R_X(\tau)| \le R_X(0)$. In other words, the maximum value of $|R_X(\tau)|$ occurs at $\tau = 0$. In particular, taking $\tau = 0$ reaffirms that $R_X(0) \ge 0$.
(*iii*) The power spectral density, $S_X(f)$, is a real, even, nonnegative function of f.

Property (i) is a restatement of the fact any correlation function, as a function of two variables, is symmetric; for a WSS process this means that $R_X(t_1 \Leftrightarrow t_2) = R_X(t_2 \Leftrightarrow t_1)$. Taking $t_1 = \tau$ and $t_2 = 0$ yields the result.

Property (ii) is a restatement of (6.1) for a WSS process. However, we can also derive it directly by again using the Cauchy–Schwarz inequality. Write

$$\begin{aligned} \big|R_X(\tau)\big| &= \big|\mathsf{E}[X_{t+\tau}\,X_t]\big| \\ &\le \sqrt{\mathsf{E}[X_{t+\tau}^2]\,\mathsf{E}[X_t^2]} \\ &= \sqrt{R_X(0)\,R_X(0)} \\ &= R_X(0). \end{aligned}$$

Property (iii): To see that $S_X(f)$ is real and even, write

$$\begin{aligned} S_X(f) &= \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau}\,dt \\ &= \int_{-\infty}^{\infty} R_X(\tau)\cos(2\pi f\tau)\,d\tau \Leftrightarrow j\int_{-\infty}^{\infty} R_X(\tau)\sin(2\pi f\tau)\,d\tau. \end{aligned}$$

Since correlation functions are real and even, and since the sine is odd, the second integrand is odd, and therefore integrates to zero. Hence, we can always write

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau)\cos(2\pi f\tau)\,d\tau.$$

Thus, $S_X(f)$ is real. Furthermore, since the cosine is even, $S_X(f)$ is an even function of f as well.

The fact that $S_X(f)$ is nonnegative is derived later in Example 6.9.

Remark. Property (iii) actually implies Properties (i) and (ii). (Problem 9.) Hence, a real-valued function of τ is a correlation function if and only if its Fourier transform is real, even, and nonnegative. However, if one is trying to show that a function of τ is *not* a correlation function, it is easier to check if either (i) or (ii) fails. Of course, if (i) and (ii) both hold, it is then necessary to go ahead and find the power spectral density and see if it is nonnegative.

***Example* 6.6.** Determine whether or not $R(\tau) := \tau e^{-|\tau|}$ is a valid correlation function.

Solution. Property (i) requires that $R(\tau)$ be even. However, $R(\tau)$ is not even (in fact, it is odd). Hence, it cannot be a correlation function.

***Example* 6.7.** Determine whether or not $R(\tau) := 1/(1+\tau^2)$ is a valid correlation function.

Solution. We must check the three properties that characterize correlation functions. It is obvious that R is even and that $\big|R(\tau)\big| \le R(0) = 1$. The Fourier

transform of $R(\tau)$ is $S(f) = \pi \exp(\Leftrightarrow 2\pi|f|)$, as can be verified by applying the inversion formula to $S(f)$.[§] Since $S(f)$ is nonnegative, we see that $R(\tau)$ satisfies all three properties, and is therefore a valid correlation function.

***Example* 6.8** (Power in a Frequency Band). Let X_t have power spectral density $S_X(f)$. Find the power in the frequency band $W_1 \leq f \leq W_2$.

Solution. Since power spectral densities are even, we include the corresponding negative frequencies too. Thus, we compute

$$\int_{-W_2}^{-W_1} S_X(f)\, df + \int_{W_1}^{W_2} S_X(f)\, df \;=\; 2\int_{W_1}^{W_2} S_X(f)\, df.$$

6.3. WSS Processes through Linear Time-Invariant Systems

In this section, we consider passing a WSS random process through a **linear time-invariant** (LTI) **system**. Our goal is to find the correlation and power spectral density of the output. Before doing so, it is convenient to introduce the following concepts.

Let X_t and Y_t be random processes. Their **cross-correlation function** is

$$R_{XY}(t_1, t_2) \;:=\; \mathsf{E}[X_{t_1} Y_{t_2}].$$

To distinguish between the terms cross-correlation function and correlation function, the latter is sometimes referred to as the **auto-correlation function**. The **cross-covariance function** is

$$C_{XY}(t_1, t_2) := \mathsf{E}[\{X_{t_1} \Leftrightarrow m_X(t_1)\}\{Y_{t_2} \Leftrightarrow m_Y(t_2)\}] = R_{XY}(t_1, t_2) \Leftrightarrow m_X(t_1)\, m_Y(t_2).$$

A pair of processes X_t and Y_t is called **jointly wide-sense stationary** (J-WSS) if all three of the following properties hold:

- (*i*) X_t is WSS.
- (*ii*) Y_t is WSS.
- (*iii*) the cross-correlation $R_{XY}(t_1, t_2) = \mathsf{E}[X_{t_1} Y_{t_2}]$ depends on t_1 and t_2 only through the difference $t_1 \Leftrightarrow t_2$. If this is the case, we write $R_{XY}(t_1 \Leftrightarrow t_2)$ instead of $R_{XY}(t_1, t_2)$.

Remark. In checking to see if a pair of processes is J-WSS, it is usually easier to check property (*iii*) before (*ii*).

The most common situation in which we have a pair of J-WSS processes occurs when X_t is the input to an LTI system, and Y_t is the output. Recall

[§] There is a short table of transform pairs in the Problems section and inside the front cover.

that an LTI system is expressed by the convolution

$$Y_t \;=\; \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau) X_\tau \, d\tau,$$

where h is the impulse response of the system. An equivalent formula is obtained by making the change of variable $\theta = t \Leftrightarrow \tau$, $d\theta = \Leftrightarrow d\tau$. This allows us to write

$$Y_t \;=\; \int_{-\infty}^{\infty} h(\theta) X_{t-\theta} \, d\theta,$$

which we use in most of the calculations below. We now show that if the input X_t is WSS, then the output Y_t is WSS and X_t and Y_t are J-WSS.

To begin, we show that $\mathsf{E}[Y_t]$ does not depend on t. To do this, write

$$\mathsf{E}[Y_t] \;=\; \mathsf{E}\bigg[\int_{-\infty}^{\infty} h(\theta) X_{t-\theta} \, d\theta \bigg] \;=\; \int_{-\infty}^{\infty} \mathsf{E}[h(\theta) X_{t-\theta}] \, d\theta.$$

To justify bringing the expectation inside the integral, write the integral as a Riemann sum and use the linearity of expectation. More explicitly,

$$\begin{aligned}
\mathsf{E}\bigg[\int_{-\infty}^{\infty} h(\theta) X_{t-\theta} \, d\theta \bigg] &\approx \mathsf{E}\bigg[\sum_i h(\theta_i) X_{t-\theta_i} \, \Delta\theta_i \bigg] \\
&= \sum_i \mathsf{E}[h(\theta_i) X_{t-\theta_i} \, \Delta\theta_i] \\
&= \sum_i \mathsf{E}[h(\theta_i) X_{t-\theta_i}] \, \Delta\theta_i \\
&\approx \int_{-\infty}^{\infty} \mathsf{E}[h(\theta) X_{t-\theta}] \, d\theta.
\end{aligned}$$

In computing the expectation inside the integral, observe that $X_{t-\theta}$ is random, but $h(\theta)$ is not. Hence, $h(\theta)$ is a constant and can be pulled out of the expectation; i.e., $\mathsf{E}[h(\theta) X_{t-\theta}] = h(\theta) \mathsf{E}[X_{t-\theta}]$. Next, since X_t is WSS, $\mathsf{E}[X_{t-\theta}]$ does not depend on $t \Leftrightarrow \theta$, and is therefore equal to some constant, say m. Hence,

$$\mathsf{E}[Y_t] \;=\; \int_{-\infty}^{\infty} h(\theta) \, m \, d\theta \;=\; m \int_{-\infty}^{\infty} h(\theta) \, d\theta,$$

which does not depend on t.

Logically, the next step would be to show that $\mathsf{E}[Y_{t_1} Y_{t_2}]$ depends on t_1 and t_2 only through the difference $t_1 \Leftrightarrow t_2$. However, it is more efficient if we first obtain a formula for the cross-correlation, $\mathsf{E}[X_{t_1} Y_{t_2}]$, and show that it depends only on $t_1 \Leftrightarrow t_2$. Write

$$\mathsf{E}[X_{t_1} Y_{t_2}] \;=\; \mathsf{E}\bigg[X_{t_1} \int_{-\infty}^{\infty} h(\theta) X_{t_2-\theta} \, d\theta \bigg]$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} h(\theta)\mathsf{E}[X_{t_1}X_{t_2-\theta}]\,d\theta \\ &= \int_{-\infty}^{\infty} h(\theta)R_X(t_1 \Leftrightarrow [t_2 \Leftrightarrow \theta])\,d\theta \\ &= \int_{-\infty}^{\infty} h(\theta)R_X([t_1 \Leftrightarrow t_2] + \theta)\,d\theta, \end{aligned}$$

which depends only on $t_1 \Leftrightarrow t_2$ as required. We can now write

$$R_{XY}(\tau) = \int_{-\infty}^{\infty} h(\theta)R_X(\tau + \theta)\,d\theta. \tag{6.3}$$

If we make the change of variable $\beta = \Leftrightarrow\theta$, $d\beta = \Leftrightarrow d\theta$, then

$$R_{XY}(\tau) = \int_{-\infty}^{\infty} h(\Leftrightarrow\beta)R_X(\tau \Leftrightarrow \beta)\,d\beta, \tag{6.4}$$

and we see that R_{XY} is the convolution of $h(\Leftrightarrow\beta)$ with $R_X(\beta)$. This suggests that we define the **cross power spectral density** by

$$S_{XY}(f) := \int_{-\infty}^{\infty} R_{XY}(\tau)e^{-j2\pi f\tau}\,d\tau.$$

Since we are considering real-valued random variables here, we must assume $h(\theta)$ is real valued. Letting $H(f)$ denote the Fourier transform of $h(\theta)$, it follows that the Fourier transform of $h(\Leftrightarrow\theta)$ is $H(f)^*$, where the superscript $*$ denotes the complex conjugate. Thus, the Fourier transform of (6.4) is

$$S_{XY}(f) = H(f)^* S_X(f). \tag{6.5}$$

We now calculate the auto-correlation of Y_t and show that it depends only on $t_1 \Leftrightarrow t_2$. Write

$$\begin{aligned} \mathsf{E}[Y_{t_1}Y_{t_2}] &= \mathsf{E}\left[\left(\int_{-\infty}^{\infty} h(\theta)X_{t_1-\theta}\,d\theta\right)Y_{t_2}\right] \\ &= \int_{-\infty}^{\infty} h(\theta)\mathsf{E}[X_{t_1-\theta}Y_{t_2}]\,d\theta \\ &= \int_{-\infty}^{\infty} h(\theta)R_{XY}([t_1 \Leftrightarrow \theta] \Leftrightarrow t_2)\,d\theta \\ &= \int_{-\infty}^{\infty} h(\theta)R_{XY}([t_1 \Leftrightarrow t_2] \Leftrightarrow \theta)\,d\theta. \end{aligned}$$

Thus,

$$R_Y(\tau) = \int_{-\infty}^{\infty} h(\theta)R_{XY}(\tau \Leftrightarrow \theta)\,d\theta.$$

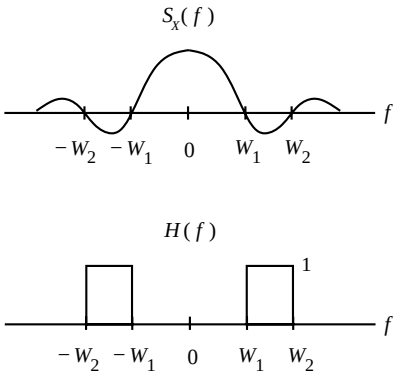

Figure 6.7. Setup of Example 6.9 to show that a power spectral density must be nonnegative.

In other words, R_Y is the convolution of h and R_{XY}. Taking Fourier transforms, we obtain $S_Y(f) = H(f)S_{XY}(f) = H(f)H(f)^* S_X(f)$, and

$$S_Y(f) \;=\; \big|H(f)\big|^2 S_X(f). \tag{6.6}$$

In other words, the power spectral density of the output of an LTI system is the product of the magnitude squared of the transfer function and the power spectral density of the input.

***Example* 6.9.** We can use the above formula to show that power spectral densities are nonnegative. Suppose X_t is a WSS process with $S_X(f) < 0$ on some interval $[W_1, W_2]$, where $0 < W_1 \le W_2$. Since S_X is even, it is also negative on $[\Leftrightarrow W_2, \Leftrightarrow W_1]$ as shown in Figure 6.7. Let $H(f) := I_{[-W_2,-W_1]}(f) + I_{[W_1,W_2]}(f)$ as shown in Figure 6.7. Note that since $H(f)$ takes only the values zero and one, $\big|H(f)\big|^2 = H(f)$. Let $h(t)$ denote the inverse Fourier transform of H. Since H is real and even, h is also real and even. Then $Y_t := \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau) X_\tau \, d\tau$ is a real WSS random process, and

$$S_Y(f) \;=\; \big|H(f)\big|^2 S_X(f) \;=\; H(f)S_X(f) \;\le\; 0,$$

with $S_Y(f) < 0$ for $W_1 < |f| < W_2$. Hence, $\int_{-\infty}^{\infty} S_Y(f)\, df < 0$. It then follows that

$$0 \;\le\; \mathsf{E}[Y_t^2] \;=\; R_Y(0) \;=\; \int_{-\infty}^{\infty} S_Y(f)\, df \;<\; 0,$$

which is a contradiction. Thus, $S_X(f) \ge 0$.

***Example* 6.10.** A certain communication receiver employs a bandpass filter to reduce white noise generated in the amplifier. Suppose that the white

noise X_t has power spectral density $\mathcal{N}_0/2$ and that the filter transfer function $H(f)$ is given in Figure 6.7. Find the expected output power from the filter.

Solution. The expected output power is obtained by integrating the power spectral density of the filter output. Denoting the filter output by Y_t,

$$P_Y \;=\; \int_{-\infty}^{\infty} S_Y(f)\,df \;=\; \int_{-\infty}^{\infty} \big|H(f)\big|^2 S_X(f)\,df.$$

Since $\big|H(f)\big|^2 S_X(f) = \mathcal{N}_0/2$ for $W_1 \le |f| \le W_2$, and is zero otherwise, $P_Y = 2(\mathcal{N}_0/2)(W_2 \Leftrightarrow W_1) = \mathcal{N}_0(W_2 \Leftrightarrow W_1)$. In other words, the expected output power is $\mathcal{N}_0$ times the bandwidth of the filter.

6.4. The Matched Filter

Consider an air-traffic control system which sends out a known radar pulse. If there are no objects in range of the radar, the system returns only a noise waveform X_t, which we model as a zero-mean, WSS random process with power spectral density $S_X(f)$. If there is an object in range, the system returns the reflected radar pulse, say $v(t)$, plus the noise X_t. We wish to design a system that decides whether received waveform is noise only, X_t, or signal plus noise, $v(t) + X_t$. As an aid to achieving this goal, we propose to take the received waveform and pass it through an LTI system with impulse response $h(t)$. If the received signal is in fact $v(t) + X_t$, then the output of the linear system is

$$\int_{-\infty}^{\infty} h(t \Leftrightarrow \tau)[v(\tau) + X_\tau]\,d\tau \;=\; v_o(t) + Y_t,$$

where

$$v_o(t) \;:=\; \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau) v(\tau)\,d\tau$$

is the output signal, and

$$Y_t \;:=\; \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau) X_\tau\,d\tau$$

is the output noise process. We will now try to find the impulse response h that maximizes the output **signal-to-noise ratio** (SNR),

$$\text{SNR} \;:=\; \frac{v_o(t_0)^2}{\mathsf{E}[Y_{t_0}^2]},$$

where $v_o(t_0)^2$ is the instantaneous output signal power at time t_0, and $\mathsf{E}[Y_{t_0}^2]$ is the expected instantaneous output noise power at time t_0. Note that since $\mathsf{E}[Y_{t_0}^2] = R_Y(0) = P_Y$, we can also write

$$\text{SNR} \;=\; \frac{v_o(t_0)^2}{P_Y}.$$

Our approach will be to obtain an upper bound on the numerator of the form $v_o(t_0)^2 \le P_Y \cdot B$, where B does not depend on the impulse response h. It will then follow that

$$\text{SNR} \;=\; \frac{v_o(t_0)^2}{P_Y} \;\le\; \frac{P_Y \cdot B}{P_Y} \;=\; B.$$

We then show how to choose the impulse response so that in fact $v_o(t_0)^2 = P_Y \cdot B$. For this choice of impulse response, we then have SNR $= B$, the maximum possible value.

We begin by analyzing the denominator in the SNR. Observe that

$$P_Y \;=\; \int_{-\infty}^{\infty} S_Y(f)\,df \;=\; \int_{-\infty}^{\infty} |H(f)|^2 S_X(f)\,df.$$

To analyze the numerator, write

$$v_o(t_0) \;=\; \int_{-\infty}^{\infty} V_o(f) e^{j2\pi f t_0}\,df \;=\; \int_{-\infty}^{\infty} H(f) V(f) e^{j2\pi f t_0}\,df,$$

where $V_o(f)$ is the Fourier transform of $v_o(t)$, and $V(f)$ is the Fourier transform of $v(t)$. Next, write

$$\begin{aligned} v_o(t_0) &= \int_{-\infty}^{\infty} H(f)\sqrt{S_X(f)} \cdot \frac{V(f)e^{j2\pi f t_0}}{\sqrt{S_X(f)}}\,df \\ &= \int_{-\infty}^{\infty} H(f)\sqrt{S_X(f)} \cdot \left[\frac{V(f)^* e^{-j2\pi f t_0}}{\sqrt{S_X(f)}}\right]^* df, \end{aligned}$$

where the asterisk denotes complex conjugation. Applying the Cauchy–Schwarz inequality (see Problem 1 and the Remark following it), we obtain the upper bound,

$$|v_o(t)|^2 \;\le\; \underbrace{\int_{-\infty}^{\infty} |H(f)|^2 S_X(f)\,df}_{=\,P_Y} \cdot \underbrace{\int_{-\infty}^{\infty} \frac{|V(f)|^2}{S_X(f)}\,df}_{=:\,B}.$$

Thus,

$$\text{SNR} \;=\; \frac{|v_o(t_0)|^2}{P_Y} \;\le\; \frac{P_Y \cdot B}{P_Y} \;=\; B.$$

Now, the Cauchy–Schwarz inequality holds with equality if and only if $H(f)\sqrt{S_X(f)}$ is a multiple of $V(f)^* e^{-j2\pi f t_0}/\sqrt{S_X(f)}$. Thus, the upper bound on the SNR will be achieved if we take $H(f)$ to solve

$$H(f)\sqrt{S_X(f)} \;=\; \alpha \frac{V(f)^* e^{-j2\pi f t_0}}{\sqrt{S_X(f)}}$$

for any complex multiplier α; i.e., we should take

$$H(f) \;=\; \alpha \frac{V(f)^* e^{-j2\pi f t_0}}{S_X(f)}.$$

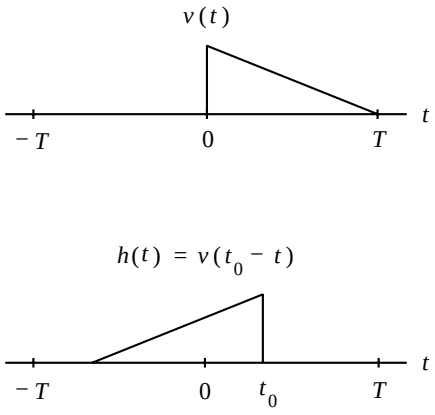

Figure 6.8. Known signal $v(t)$ and corresponding matched filter impulse response $h(t)$ in the case of white noise.

Thus, the optimal filter is "matched" to the known signal.

Consider the special case in which X_t is white noise with power spectral density $S_X(f) = \mathcal{N}_0/2$. Taking $\alpha = \mathcal{N}_0/2$ as well, we have $H(f) = V(f)^* e^{-j2\pi f t_0}$, which inverse transforms to $h(t) = v(t_0 - t)$, assuming $v(t)$ is real. Thus, the matched filter has an impulse response which is a time-reversed and translated copy of the known signal $v(t)$. An example of $v(t)$ and the corresponding $h(t)$ is shown in Figure 6.8. As the figure illustrates, if $v(t)$ is a finite-duration, causal waveform, as any radar "pulse" would be, then the sampling time t_0 can always be chosen so that $h(t)$ corresponds to a causal system.

6.5. The Wiener Filter

In the preceding section, the available data was of the form $v(t) + X_t$, where $v(t)$ was a known, nonrandom signal, and X_t was a zero-mean, WSS noise process. In this section, we suppose that V_t is an unknown random process that we would like to estimate based on observing a related random process U_t. For example, we might have $U_t = V_t + X_t$, where X_t is a noise process. However, to begin, we only assume that U_t and V_t are zero-mean, J-WSS with known power spectral densities and known cross power spectral density.

We restrict attention to linear estimators of the form

$$\widehat{V}_t \;=\; \int_{-\infty}^{\infty} h(t-\tau) U_\tau \, d\tau \;=\; \int_{-\infty}^{\infty} h(\theta) U_{t-\theta} \, d\theta. \tag{6.7}$$

Note that to estimate V_t at a single time t, we use the entire observed waveform U_τ for $-\infty < \tau < \infty$. Our goal is to find an impulse response h that minimizes the **mean squared error**, $\mathsf{E}[|V_t - \widehat{V}_t|^2]$. In other words, we are looking for an impulse response h such that if $\widehat{V}_t$ is given by (6.7), and if $\tilde{h}$ is any other

impulse response, and we put

$$\widetilde{V}_t \;=\; \int_{-\infty}^{\infty} \tilde{h}(t \Leftrightarrow \tau) U_\tau \, d\tau \;=\; \int_{-\infty}^{\infty} \tilde{h}(\theta) U_{t-\theta} \, d\theta, \tag{6.8}$$

then

$$\mathsf{E}[|V_t \Leftrightarrow \widehat{V}_t|^2] \;\le\; \mathsf{E}[|V_t \Leftrightarrow \widetilde{V}_t|^2].$$

To find the optimal filter h, we apply the **orthogonality principle**, which says that if

$$\mathsf{E}\left[(V_t \Leftrightarrow \widehat{V}_t) \int_{-\infty}^{\infty} \tilde{h}(\theta) U_{t-\theta} \, d\theta\right] \;=\; 0 \tag{6.9}$$

for *every* filter $\tilde{h}$, then h is the optimal filter.

Before proceeding any further, we need the following observation. Suppose (6.9) holds for *every* choice of $\tilde{h}$. Then in particular, it holds if we replace $\tilde{h}$ by $h \Leftrightarrow \tilde{h}$. Making this substitution in (6.9) yields

$$\mathsf{E}\left[(V_t \Leftrightarrow \widehat{V}_t) \int_{-\infty}^{\infty} [h(\theta) \Leftrightarrow \tilde{h}(\theta)] U_{t-\theta} \, d\theta\right] \;=\; 0.$$

Since the integral in this expression is simply $\widehat{V}_t \Leftrightarrow \widetilde{V}_t$, we have that

$$\mathsf{E}[(V_t \Leftrightarrow \widehat{V}_t)(\widehat{V}_t \Leftrightarrow \widetilde{V}_t)] \;=\; 0. \tag{6.10}$$

To establish the orthogonality principle, assume (6.9) holds for every choice of $\tilde{h}$. Then (6.10) holds as well. Now write

$$\begin{aligned}
\mathsf{E}[|V_t \Leftrightarrow \widetilde{V}_t|^2] &= \mathsf{E}[|(V_t \Leftrightarrow \widehat{V}_t) + (\widehat{V}_t \Leftrightarrow \widetilde{V}_t)|^2] \\
&= \mathsf{E}[|V_t \Leftrightarrow \widehat{V}_t|^2 + 2(V_t \Leftrightarrow \widehat{V}_t)(\widehat{V}_t \Leftrightarrow \widetilde{V}_t) + |\widehat{V}_t \Leftrightarrow \widetilde{V}_t|^2] \\
&= \mathsf{E}[|V_t \Leftrightarrow \widehat{V}_t|^2] + 2\mathsf{E}[(V_t \Leftrightarrow \widehat{V}_t)(\widehat{V}_t \Leftrightarrow \widetilde{V}_t)] + \mathsf{E}[|V_t \Leftrightarrow \widetilde{V}_t|^2] \\
&= \mathsf{E}[|V_t \Leftrightarrow \widehat{V}_t|^2] + \mathsf{E}[|V_t \Leftrightarrow \widetilde{V}_t|^2] \\
&\ge \mathsf{E}[|V_t \Leftrightarrow \widehat{V}_t|^2],
\end{aligned}$$

and thus, h is the filter that minimizes the mean squared error.

The next task is to characterize the filter h such that (6.9) holds for every choice of $\tilde{h}$. Recalling (6.9), we have

$$\begin{aligned}
0 &= \mathsf{E}\left[(V_t \Leftrightarrow \widehat{V}_t) \int_{-\infty}^{\infty} \tilde{h}(\theta) U_{t-\theta} \, d\theta\right] \\
&= \mathsf{E}\left[\int_{-\infty}^{\infty} \tilde{h}(\theta)(V_t \Leftrightarrow \widehat{V}_t) U_{t-\theta} \, d\theta\right] \\
&= \int_{-\infty}^{\infty} \mathsf{E}[\tilde{h}(\theta)(V_t \Leftrightarrow \widehat{V}_t) U_{t-\theta}] \, d\theta \\
&= \int_{-\infty}^{\infty} \tilde{h}(\theta) \mathsf{E}[(V_t \Leftrightarrow \widehat{V}_t) U_{t-\theta}] \, d\theta \\
&= \int_{-\infty}^{\infty} \tilde{h}(\theta) [R_{VU}(\theta) \Leftrightarrow R_{\widehat{V}U}(\theta)] \, d\theta.
\end{aligned}$$

Since $\tilde{h}$ is arbitrary, take $\tilde{h}(\theta) = R_{VU}(\theta) \Leftrightarrow R_{\widehat{V}U}(\theta)$ to get

$$\int_{-\infty}^{\infty} \left| R_{VU}(\theta) \Leftrightarrow R_{\widehat{V}U}(\theta) \right|^2 d\theta \;=\; 0. \tag{6.11}$$

Thus, (6.9) holds for all $\tilde{h}$ if and only if $R_{VU} = R_{\widehat{V}U}$.

The next task is to analyze $R_{\widehat{V}U}$. Recall that $\widehat{V}_t$ in (6.7) is the response of an LTI system to input U_t. Applying (6.3) with X replaced by U and Y replaced by $\widehat{V}$, we have, also using the fact that R_U is even,

$$R_{\widehat{V}U}(\tau) \;=\; R_{U\widehat{V}}(\Leftrightarrow\tau) \;=\; \int_{-\infty}^{\infty} h(\theta) R_U(\tau \Leftrightarrow \theta)\, d\theta.$$

Taking Fourier transforms of

$$R_{VU}(\tau) \;=\; R_{\widehat{V}U}(\tau) \;=\; \int_{-\infty}^{\infty} h(\theta) R_U(\tau \Leftrightarrow \theta)\, d\theta \tag{6.12}$$

yields

$$S_{VU}(f) \;=\; H(f) S_U(f).$$

Thus,

$$H(f) \;=\; \frac{S_{VU}(f)}{S_U(f)}$$

is the optimal filter. This choice of $H(f)$ is called the **Wiener filter**.

* *Causal Wiener Filters*

Typically, the Wiener filter as found above is not causal; i.e., we do not have $h(t) = 0$ for $t < 0$. To find such an h, we need to reformulate the problem by replacing (6.7) with

$$\widehat{V}_t \;=\; \int_{-\infty}^{t} h(t \Leftrightarrow \tau) U_\tau \, d\tau \;=\; \int_{0}^{\infty} h(\theta) U_{t-\theta} \, d\theta,$$

and replacing (6.8) with

$$\widetilde{V}_t \;=\; \int_{-\infty}^{t} \tilde{h}(t \Leftrightarrow \tau) U_\tau \, d\tau \;=\; \int_{0}^{\infty} \tilde{h}(\theta) U_{t-\theta} \, d\theta.$$

Everything proceeds as before from (6.9) through (6.11) except that lower limits of integration are changed from $\Leftrightarrow\infty$ to 0. Thus, instead of concluding $R_{VU}(\tau) = R_{\widehat{V}U}(\tau)$ for all τ, we only have $R_{VU}(\tau) = R_{\widehat{V}U}(\tau)$ for $\tau \geq 0$. Instead of (6.12), we have

$$R_{VU}(\tau) \;=\; \int_{0}^{\infty} h(\theta) R_U(\tau \Leftrightarrow \theta)\, d\theta, \quad \tau \geq 0. \tag{6.13}$$

This is known as the **Wiener–Hopf equation**. Because the equation only holds for $\tau \geq 0$, we run into a problem if we try to take Fourier transforms. To

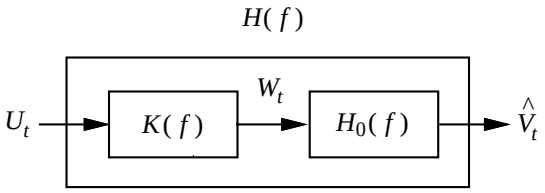

Figure 6.9. Decomposition of the causal Wiener filter using the whitening filter $K(f)$.

compute $S_{VU}(f)$, we need to integrate $R_{VU}(\tau)e^{-j2\pi f\tau}$ from $\tau = -\infty$ to $\tau = \infty$. But we can only use the Wiener–Hopf equation for $\tau \geq 0$.

In general, the Wiener–Hopf equation is very difficult to solve. However, if U is white noise, say $R_U(\theta) = \delta(\theta)$, then (6.13) reduces to

$$R_{VU}(\tau) \;=\; h(\tau), \quad \tau \geq 0.$$

Since h is causal, $h(\tau) = 0$ for $\tau < 0$.

The preceding observation suggests the construction of $H(f)$ using a **whitening filter** as shown in Figure 6.9. If U_t is not white noise, suppose we can find a causal filter $K(f)$ such that when U_t is passed through this system, the output is white noise W_t, by which we mean $S_W(f) = 1$. Letting k denote the impulse response corresponding to K, we can write W_t mathematically as

$$W_t \;=\; \int_0^\infty k(\theta)U_{t-\theta}\,d\theta. \tag{6.14}$$

Then

$$1 \;=\; S_W(f) \;=\; |K(f)|^2 S_U(f). \tag{6.15}$$

Consider the problem of causally estimating V_t based on W_t instead of U_t. The solution is again given by the Wiener–Hopf equation,

$$R_{VW}(\tau) \;=\; \int_0^\infty h_0(\theta)R_W(\tau - \theta)\,d\theta, \quad \tau \geq 0.$$

Since K was chosen so that $S_W(f) = 1$, $R_W(\theta) = \delta(\theta)$. Therefore, the Wiener–Hopf equation tells us that $h_0(\tau) = R_{VW}(\tau)$ for $\tau \geq 0$. Using (6.14), it is easy to see that

$$R_{VW}(\tau) \;=\; \int_0^\infty k(\theta)R_{VU}(\tau + \theta)\,d\theta, \tag{6.16}$$

and then[¶]

$$S_{VW}(f) \;=\; K(f)^* S_{VU}(f). \tag{6.17}$$

We now summarize the procedure.

[¶] If $k(\theta)$ is complex valued, so is W_t in (6.14). In this case, as in Problem 26, it is understood that $R_{VW}(\tau) = \mathsf{E}[V_{t+\tau}W_t^*]$.

1. According to (6.15), we must first write $S_U(f)$ in the form

$$S_U(f) \;=\; \frac{1}{K(f)} \cdot \frac{1}{K(f)^*},$$

where $K(f)$ is a causal filter (this is known as **spectral factorization**).‖

2. The optimum filter is $H(f) = H_0(f)K(f)$, where

$$H_0(f) \;=\; \int_0^\infty R_{VW}(\tau) e^{-j\,2\pi f \tau}\, d\tau,$$

and $R_{VW}(\tau)$ is given by (6.16) or by the inverse transform of (6.17).

***Example* 6.11.** Let $U_t = V_t + X_t$, where V_t and X_t are zero-mean, WSS processes with $\mathsf{E}[V_t X_\tau] = 0$ for all t and τ. Assume that the signal V_t has power spectral density $S_V(f) = 2\lambda/[\lambda^2 + (2\pi f)^2]$ and that the noise X_t is white with power spectral density $S_X(f) = 1$. Find the causal Wiener filter.

Solution. From your solution of Problem 32, $S_U(f) = S_V(f) + S_X(f)$. Thus,

$$S_U(f) \;=\; \frac{2\lambda}{\lambda^2 + (2\pi f)^2} + 1 \;=\; \frac{A^2 + (2\pi f)^2}{\lambda^2 + (2\pi f)^2},$$

where $A^2 := \lambda^2 + 2\lambda$. This factors into

$$S_U(f) \;=\; \frac{A + j2\pi f}{\lambda + j2\pi f} \cdot \frac{A \Leftrightarrow j2\pi f}{\lambda \Leftrightarrow j2\pi f}.$$

Then

$$K(f) \;=\; \frac{\lambda + j2\pi f}{A + j2\pi f}$$

is the required causal (by Problem 35) whitening filter. Next, from your solution of Problem 32, $S_{VU}(f) = S_V(f)$. So, by (6.17),

$$\begin{aligned} S_{VW}(f) &= \frac{\lambda \Leftrightarrow j2\pi f}{A \Leftrightarrow j2\pi f} \cdot \frac{2\lambda}{\lambda^2 + (2\pi f)^2} \\ &= \frac{2\lambda}{(A \Leftrightarrow j2\pi f)(\lambda + j2\pi f)} \\ &= \frac{B}{A \Leftrightarrow j2\pi f} + \frac{B}{\lambda + j2\pi f}, \end{aligned}$$

where $B := 2\lambda/(\lambda + A)$. It follows that

$$R_{VW}(\tau) \;=\; Be^{A\tau}\, u(\Leftrightarrow\tau) + Be^{-\lambda\tau} u(\tau),$$

‖If $S_U(f)$ satisfies the **Paley–Wiener condition**,

$$\int_{-\infty}^{\infty} \frac{|\ln S_U(f)|}{1+f^2}\, df \;<\; \infty,$$

then $S_U(f)$ can always be factored in this way.

where u is the unit-step function. Since $h_0(\tau) = R_{VW}(\tau)$ for $\tau \geq 0$, $h_0(\tau) = Be^{-\lambda\tau}u(\tau)$ and $H_0(f) = B/(\lambda + j2\pi f)$. Next,

$$H(f) \;=\; H_0(f)K(f) \;=\; \frac{B}{\lambda + j2\pi f} \cdot \frac{\lambda + j2\pi f}{A + j2\pi f} \;=\; \frac{B}{A + j2\pi f},$$

and $h(\tau) = Be^{-A\tau}u(\tau)$.

6.6. *Expected Time-Average Power and the Wiener–Khinchin Theorem

In this section we develop the notion of the expected time-average power in a WSS process. We then derive the Wiener–Khinchin Theorem, which implies that the expected time-average power is the same as the expected instantaneous power. As an easy corollary of the derivation of the Wiener–Khinchin Theorem, we derive the mean-square ergodic theorem for WSS processes. This result shows that $\mathsf{E}[X_t]$ can often be computed by averaging a single sample path over time.

Recall that the average power in a deterministic waveform $x(t)$ is given by

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)^2 \, dt.$$

The analogous formula for a random process is

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X_t^2 \, dt.$$

The problem with the above quantity is that it is a random variable because the integrand, X_t^2, is random. We would like to characterize the process by a nonrandom quantity. This suggests that we define the **expected time-average power** in a random process by

$$\widetilde{P}_X \;:=\; \mathsf{E}\left[\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X_t^2 \, dt\right].$$

We now carry out some analysis of $\widetilde{P}_X$. To begin, we want to apply **Parseval's equation** to the integral. To do this, we use the following notation. Let

$$X_t^T \;:=\; \begin{cases} X_t, & |t| \leq T, \\ 0, & |t| > T, \end{cases}$$

so that $\int_{-T}^{T} X_t^2 \, dt = \int_{-\infty}^{\infty} \left|X_t^T\right|^2 dt$. Now, the Fourier transform of X_t^T is

$$\widetilde{X}_f^T \;:=\; \int_{-\infty}^{\infty} X_t^T e^{-j2\pi ft} \, dt \;=\; \int_{-T}^{T} X_t \, e^{-j2\pi ft} \, dt, \tag{6.18}$$

and by Parseval's equation,

$$\int_{-\infty}^{\infty} \left| X_t^T \right|^2 dt \;=\; \int_{-\infty}^{\infty} \left| \widetilde{X}_f^T \right|^2 df.$$

Returning now to $\widetilde{P}_X$, we have

$$\begin{aligned}
\widetilde{P}_X &= \mathsf{E}\left[\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X_t^2 \, dt\right] \\
&= \mathsf{E}\left[\lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} \left| X_t^T \right|^2 dt\right] \\
&= \mathsf{E}\left[\lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} \left| \widetilde{X}_f^T \right|^2 df\right] \\
&= \int_{-\infty}^{\infty} \left(\lim_{T\to\infty} \frac{\mathsf{E}\left[\left| \widetilde{X}_f^T \right|^2\right]}{2T}\right) df.
\end{aligned}$$

The **Wiener–Khinchin Theorem** says that the above integrand is exactly the power spectral density $S_X(f)$. From our earlier work, we know that $P_X := \mathsf{E}[X_t^2] = R_X(0) = \int_{-\infty}^{\infty} S_X(f)\, df$. Thus, $\widetilde{P}_X = P_X$; i.e., the expected time-average power is equal to the expected instantaneous power at any point in time. Note also that the above integrand is obviously nonnegative, and so if we can show that it is equal to $S_X(f)$, then this will provide another proof that $S_X(f)$ must be nonnegative.

To derive the Wiener–Khinchin Theorem, we begin with the numerator,

$$\mathsf{E}\left[\left| \widetilde{X}_f^T \right|^2\right] \;=\; \mathsf{E}\left[\left(\widetilde{X}_f^T\right)\left(\widetilde{X}_f^T\right)^*\right],$$

where the asterisk denotes complex conjugation. To evaluate the right-hand side, use (6.18) to obtain

$$\mathsf{E}\left[\left(\int_{-T}^{T} X_t e^{-j\,2\pi f t}\, dt\right)\left(\int_{-T}^{T} X_\theta\, e^{-j\,2\pi f \theta}\, d\theta\right)^*\right].$$

We can now write

$$\begin{aligned}
\mathsf{E}\left[\left| \widetilde{X}_f^T \right|^2\right] &= \int_{-T}^{T}\int_{-T}^{T} \mathsf{E}[X_t X_\theta] e^{-j\,2\pi f(t-\theta)}\, dt\, d\theta \\
&= \int_{-T}^{T}\int_{-T}^{T} R_X(t \Leftrightarrow \theta) e^{-j\,2\pi f(t-\theta)}\, dt\, d\theta \qquad (6.19) \\
&= \int_{-T}^{T}\int_{-T}^{T} \left[\int_{-\infty}^{\infty} S_X(\nu) e^{j\,2\pi\nu(t-\theta)}\, d\nu\right] e^{-j\,2\pi f(t-\theta)}\, dt\, d\theta \\
&= \int_{-\infty}^{\infty} S_X(\nu) \left[\int_{-T}^{T} e^{j\,2\pi\theta(f-\nu)} \left(\int_{-T}^{T} e^{j\,2\pi t(\nu-f)}\, dt\right) d\theta\right] d\nu.
\end{aligned}$$

Notice that the inner two integrals decouple so that

$$\begin{aligned}\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big] &= \int_{-\infty}^{\infty} S_X(\nu)\left[\int_{-T}^{T} e^{j2\pi\theta(f-\nu)}\,d\theta\right]\left(\int_{-T}^{T} e^{j2\pi t(\nu-f)}\,dt\right)d\nu \\ &= \int_{-\infty}^{\infty} S_X(\nu)\cdot\left[2T\frac{\sin\big(2\pi T(f\Leftrightarrow\nu)\big)}{2\pi T(f\Leftrightarrow\nu)}\right]^2 d\nu.\end{aligned}$$

We can then write

$$\frac{\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big]}{2T} = \int_{-\infty}^{\infty} S_X(\nu)\cdot 2T\left[\frac{\sin\big(2\pi T(f\Leftrightarrow\nu)\big)}{2\pi T(f\Leftrightarrow\nu)}\right]^2 d\nu. \tag{6.20}$$

This is a convolution integral. Furthermore, the quantity multiplying $S_X(\nu)$ converges to the delta function $\delta(f\Leftrightarrow\nu)$ as $T\to\infty$.[2] Thus,

$$\lim_{T\to\infty}\frac{\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big]}{2T} = \int_{-\infty}^{\infty} S_X(\nu)\,\delta(f\Leftrightarrow\nu)\,d\nu = S_X(f),$$

which is exactly the Wiener–Khinchin Theorem.

Remark. The preceding derivation shows that $S_X(f)$ is equal to the limit of (6.19) divided by $2T$. Thus,

$$S(f) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} R_X(t\Leftrightarrow\theta)e^{-j2\pi f(t-\theta)}\,dt\,d\theta.$$

As noted in Problem 37, the properties of the correlation function directly imply that this double integral is nonnegative. This is the direct way to prove that power spectral densities are nonnegative.

Mean-Square Law of Large Numbers for WSS Processes

In the process of deriving the weak law of large numbers in Chapter 2, we showed that for an uncorrelated sequence X_n with common mean $m=\mathsf{E}[X_n]$ and common variance $\sigma^2=\mathsf{var}(X_n)$, the sample mean (or time average)

$$M_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges to m in the sense that $\mathsf{E}[|M_n\Leftrightarrow m|^2]=\mathsf{var}(M_n)\to 0$ as $n\to\infty$ by (2.6). In this case, we say that M_n converges in mean square to m, and we call this a **mean-square law of large numbers**.

We now show that for a WSS process Y_t with mean $m=\mathsf{E}[Y_t]$, the sample mean (or time average)

$$M_T := \frac{1}{2T}\int_{-T}^{T} Y_t\,dt \to m$$

in the sense that $\mathsf{E}[|M_T \Leftrightarrow m|^2] \to 0$ as $T \to \infty$ if and only if

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-2T}^{2T} C_Y(\tau)\Big(1 \Leftrightarrow \frac{|\tau|}{2T}\Big)\, d\tau \;=\; 0,$$

where C_Y is the covariance (not correlation) function of Y_t. We can view this result as a mean-square law of large numbers for WSS processes. Laws of large numbers for sequences or processes that are not uncorrelated are often called **ergodic theorems**. Hence, the above result is also called the **mean-square ergodic theorem for WSS processes**. The point in all theorems of this type is that the expectation $\mathsf{E}[Y_t]$ can be computed by averaging a single sample path over time.

We now derive this result. To begin, put $X_t := Y_t \Leftrightarrow m$ so that X_t is zero mean and has correlation function $R_X(\tau) = C_Y(\tau)$. Also,

$$M_T \Leftrightarrow m \;=\; \frac{1}{2T}\int_{-T}^{T} X_t\, dt \;=\; \frac{\widetilde{X}_f^T\big|_{f=0}}{2T},$$

where $\widetilde{X}_f^T$ was defined in (6.18). Using (6.20) with $f = 0$ we can write

$$\begin{aligned}
\mathsf{E}[|M_T \Leftrightarrow m|^2] &= \frac{\mathsf{E}\big[\big|\widetilde{X}_0^T\big|^2\big]}{4T^2} \\
&= \frac{1}{2T}\cdot\frac{\mathsf{E}\big[\big|\widetilde{X}_0^T\big|^2\big]}{2T} \\
&= \frac{1}{2T}\int_{-\infty}^{\infty} S_X(\nu)\cdot 2T\left[\frac{\sin(2\pi T\nu)}{2\pi T\nu}\right]^2 d\nu. \qquad (6.21)
\end{aligned}$$

Applying Parseval's equation yields

$$\begin{aligned}
\mathsf{E}[|M_T \Leftrightarrow m|^2] &= \frac{1}{2T}\int_{-2T}^{2T} R_X(\tau)\Big(1 \Leftrightarrow \frac{|\tau|}{2T}\Big)\, d\tau \\
&= \frac{1}{2T}\int_{-2T}^{2T} C_Y(\tau)\Big(1 \Leftrightarrow \frac{|\tau|}{2T}\Big)\, d\tau.
\end{aligned}$$

To give a sufficient condition for when this goes to zero, recall that by the argument following (6.20), as $T \to \infty$ the integral in (6.21) is approximately $S_X(0)$ if $S_X(f)$ is continuous at $f = 0$.[3] Thus, if $S_X(f)$ is continuous at $f = 0$,

$$\mathsf{E}[|M_T \Leftrightarrow m|^2] \;\approx\; \frac{S_X(0)}{2T} \;\to\; 0$$

as $T \to \infty$.

Remark. If $R_X(\tau) = C_Y(\tau)$ is absolutely integrable, then $S_X(f)$ is uniformly continuous. To see this write

$$|S_X(f) \Leftrightarrow S_X(f_0)| \;=\; \left|\int_{-\infty}^{\infty} R_X(\tau)e^{-j\,2\pi f\tau}\, d\tau \Leftrightarrow \int_{-\infty}^{\infty} R_X(\tau)e^{-j\,2\pi f_0\tau}\, d\tau\right|$$

$$\begin{aligned} &\leq \int_{-\infty}^{\infty} |R_X(\tau)|\,|e^{-j2\pi f\tau} \Leftrightarrow e^{-j2\pi f_0\tau}|\,d\tau \\ &= \int_{-\infty}^{\infty} |R_X(\tau)|\,|e^{-j2\pi f_0\tau}[e^{-j2\pi(f-f_0)\tau} \Leftrightarrow 1]|\,d\tau \\ &= \int_{-\infty}^{\infty} |R_X(\tau)|\,|e^{-j2\pi(f-f_0)\tau} \Leftrightarrow 1|\,d\tau. \end{aligned}$$

Now observe that $|e^{-j2\pi(f-f_0)\tau} \Leftrightarrow 1| \to 0$ as $f \to f_0$. Since R_X is absolutely integrable, Lebesgue's dominated convergence theorem [4, p. 209] implies that the integral goes to zero as well.

6.7. ⋆Power Spectral Densities for non-WSS Processes

If X_t is not WSS, then the instantaneous power $\mathsf{E}[X_t^2]$ will depend on t. In this case, it is more appropriate to look at the expected time-average power $\widetilde{P}_X$ defined in the previous section. At the end of this section, we show that

$$\lim_{T\to\infty} \frac{\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big]}{2T} = \int_{-\infty}^{\infty} \overline{R}_X(\tau) e^{-j2\pi f\tau}\,d\tau, \tag{6.22}$$

where**

$$\overline{R}_X(\tau) := \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} R_X(\tau+\theta,\theta)\,d\theta.$$

Hence, for a non-WSS process, we define its power spectral density to be the Fourier transform of $\overline{R}_X(\tau)$,

$$S_X(f) := \int_{-\infty}^{\infty} \overline{R}_X(\tau) e^{-j2\pi f\tau}\,d\tau.$$

An important application the foregoing is to **cyclostationary** processes. A process Y_t is (wide-sense) cyclostationary if its mean function is periodic in t, and if its correlation function has the property that for fixed τ, $R_X(\tau+\theta,\theta)$ is periodic in θ. For a cyclostationary process with period T_0, it is not hard to show that

$$\overline{R}_X(\tau) = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} R_X(\tau+\theta,\theta)\,d\theta. \tag{6.23}$$

***Example* 6.12.** Let X_t be WSS, and put $Y_t := X_t\cos(2\pi f_0 t)$. Show that Y_t is cyclostationary and that

$$S_Y(f) = \frac{1}{4}[S_X(f \Leftrightarrow f_0) + S_X(f+f_0)].$$

Solution. The mean of Y_t is

$$\mathsf{E}[Y_t] = \mathsf{E}[X_t\cos(2\pi f_0 t)] = \mathsf{E}[X_t]\cos(2\pi f_0 t).$$

**Note that if X_t is WSS, then $\overline{R}_X(\tau) = R_X(\tau)$.

Because X_t is WSS, $\mathsf{E}[X_t]$ does not depend on t, and it is then clear that $\mathsf{E}[Y_t]$ has period $1/f_0$. Next consider

$$\begin{aligned} R_Y(t+\theta,\theta) &= \mathsf{E}[Y_{t+\theta}Y_\theta] \\ &= \mathsf{E}[X_{t+\theta}\cos(2\pi f_0\{t+\theta\})X_\theta\cos(2\pi f_0\theta)] \\ &= R_X(t)\cos(2\pi f_0\{t+\theta\})\cos(2\pi f_0\theta), \end{aligned}$$

which is periodic in θ with period $1/f_0$. To compute $S_Y(f)$, first use a trigonometric identity to write

$$R_Y(t+\theta,\theta) = \frac{R_X(t)}{2}[\cos(2\pi f_0 t)+\cos(2\pi f_0\{t+2\theta\})].$$

Applying (6.23) to R_Y with $T_0 = 1/f_0$ yields

$$\overline{R}_Y(t) = \frac{R_X(t)}{2}\cos(2\pi f_0 t).$$

Taking Fourier transforms yields the claimed formula for $S_Y(f)$.

Derivation of (6.22)

We begin as in the derivation of the Wiener–Khinchin Theorem, except that instead of (6.19) we have

$$\begin{aligned} \mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big] &= \int_{-T}^{T}\int_{-T}^{T} R_X(t,\theta)e^{-j2\pi f(t-\theta)}\,dt\,d\theta \\ &= \int_{-T}^{T}\int_{-\infty}^{\infty} I_{[-T,T]}(t)R_X(t,\theta)e^{-j2\pi f(t-\theta)}\,dt\,d\theta. \end{aligned}$$

Now make the change of variable $\tau = t \Leftrightarrow \theta$ in the inner integral. This results in

$$\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big] = \int_{-T}^{T}\int_{-\infty}^{\infty} I_{[-T,T]}(\tau+\theta)R_X(\tau+\theta,\theta)e^{-j2\pi f\tau}\,d\tau\,d\theta.$$

Change the order of integration to get

$$\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big] = \int_{-\infty}^{\infty} e^{-j2\pi f\tau}\int_{-T}^{T} I_{[-T,T]}(\tau+\theta)R_X(\tau+\theta,\theta)\,d\theta\,d\tau.$$

To simplify the inner integral, observe that $I_{[-T,T]}(\tau+\theta) = I_{[-T-\tau,T-\tau]}(\theta)$. Now $T \Leftrightarrow \tau$ is to the left of $\Leftrightarrow T$ if $2T < \tau$, and $\Leftrightarrow T \Leftrightarrow \tau$ is to the right of T if $\Leftrightarrow 2T > \tau$. Thus,

$$\frac{\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big]}{2T} = \int_{-\infty}^{\infty} e^{-j2\pi f\tau}\,g_T(\tau)\,d\tau,$$

where

$$g_T(\tau) \;:=\; \begin{cases} \frac{1}{2T}\displaystyle\int_{-T}^{T-\tau} R_X(\tau+\theta,\theta)\,d\theta, & 0 \le \tau \le 2T, \\ \frac{1}{2T}\displaystyle\int_{-T-\tau}^{T} R_X(\tau+\theta,\theta)\,d\theta, & \Leftarrow 2T \le \tau < 0, \\ 0, & |\tau| > 2T. \end{cases}$$

If T much greater than $|\tau|$, then $T \Leftarrow \tau \approx T$ and $\Leftarrow T \Leftarrow \tau \approx \Leftarrow T$ in the above limits of integration. Hence, if R_X is a reasonably-behaved correlation function,

$$\lim_{T\to\infty} g_T(\tau) \;=\; \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} R_X(\tau+\theta,\theta)\,d\theta \;=\; \overline{R}_X(\tau),$$

and we find that

$$\lim_{T\to\infty} \frac{\mathsf{E}\big[\big|\widetilde{X}_f^T\big|^2\big]}{2T} \;=\; \int_{-\infty}^{\infty} e^{-j2\pi f\tau} \lim_{T\to\infty} g_T(\tau)\,d\tau \;=\; \int_{-\infty}^{\infty} e^{-j2\pi f\tau}\,\overline{R}_X(\tau)\,d\tau.$$

Remark. If X_t is actually WSS, then $R_X(\tau+\theta,\theta) = R_X(\tau)$, and

$$g_T(\tau) \;=\; R_X(\tau)\Big(1 \Leftarrow \frac{|\tau|}{2T}\Big) I_{[-2T,2T]}(\tau).$$

In this case, for each fixed τ, $g_T(\tau) \to R_X(\tau)$. We thus have an alternative derivation of the Wiener–Khinchin Theorem.

6.8. Notes

Notes §6.2: Wide-Sense Stationary Processes

***Note* 1.** We mean that if R_X is the correlation function of a WSS process, then R_X it must satisfy Properties (i)–(iii). Furthermore, it can be proved [50, pp. 94–96] that if R is a real-valued function satisfying Properties (i)–(iii), then there is a WSS process having R as its correlation function.

Notes §6.6: *Expected Time-Average Power and the Wiener–Khinchin Theorem

***Note* 2.** To give a rigorous derivation of the fact that

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} S_X(\nu)\cdot 2T\left[\frac{\sin\big(2\pi T(f \Leftarrow \nu)\big)}{2\pi T(f \Leftarrow \nu)}\right]^2 d\nu \;=\; S_X(f),$$

it is convenient to assume $S_X(f)$ is continuous at f. Letting

$$\delta_T(f) \;:=\; 2T\left[\frac{\sin(2\pi T f)}{2\pi T f}\right]^2,$$

we must show that

$$\left| \int_{-\infty}^{\infty} S_X(\nu)\, \delta_T(f \Leftrightarrow \nu)\, d\nu \Leftrightarrow S_X(f) \right| \;\rightarrow\; 0.$$

To proceed, we need the following properties of δ_T. First,

$$\int_{-\infty}^{\infty} \delta_T(f)\, df \;=\; 1.$$

This can be seen by using the Fourier transform table to evaluate the inverse transform of $\delta_T(f)$ at $t = 0$. Second, for fixed $\Delta f > 0$, as $T \rightarrow \infty$,

$$\int_{\{f:|f|>\Delta f\}} \delta_T(f)\, df \;\rightarrow\; 0.$$

This can be seen by using the fact that $\delta_T(f)$ is even and writing

$$\int_{\Delta f}^{\infty} \delta_T(f)\, df \;\leq\; \frac{2T}{(2\pi T)^2} \int_{\Delta f}^{\infty} \frac{1}{f^2}\, df \;=\; \frac{1}{2T\pi^2 \Delta f},$$

which goes to zero as $T \rightarrow \infty$. Third, for $|f| \geq \Delta f > 0$,

$$|\delta_T(f)| \;\leq\; \frac{1}{2T(\pi \Delta f)^2}.$$

Now, using the first property of δ_T, write

$$S_X(f) \;=\; S_X(f) \int_{-\infty}^{\infty} \delta_T(f \Leftrightarrow \nu)\, d\nu \;=\; \int_{-\infty}^{\infty} S_X(f)\, \delta_T(f \Leftrightarrow \nu)\, d\nu.$$

Then

$$S_X(f) \Leftrightarrow \int_{-\infty}^{\infty} S_X(\nu)\, \delta_T(f \Leftrightarrow \nu)\, d\nu \;=\; \int_{-\infty}^{\infty} [S_X(f) \Leftrightarrow S_X(\nu)]\, \delta_T(f \Leftrightarrow \nu)\, d\nu.$$

For the next step, let $\varepsilon > 0$ be given, and use the continuity of S_X at f to get the existence of a $\Delta f > 0$ such that for $|f \Leftrightarrow \nu| < \Delta f$, $|S_X(f) \Leftrightarrow S_X(\nu)| < \varepsilon$. Now break up the range of integration into ν such that $|f \Leftrightarrow \nu| < \Delta f$ and ν such that $|f \Leftrightarrow \nu| \geq \Delta f$. For the first range, we need the calculation

$$\begin{aligned}
&\left| \int_{f-\Delta f}^{f+\Delta f} [S_X(f) \Leftrightarrow S_X(\nu)]\, \delta_T(f \Leftrightarrow \nu)\, d\nu \right| \\
&\qquad \leq \int_{f-\Delta f}^{f+\Delta f} |S_X(f) \Leftrightarrow S_X(\nu)|\, \delta_T(f \Leftrightarrow \nu)\, d\nu \\
&\qquad \leq \varepsilon \int_{f-\Delta f}^{f+\Delta f} \delta_T(f \Leftrightarrow \nu)\, d\nu \\
&\qquad \leq \varepsilon \int_{-\infty}^{\infty} \delta_T(f \Leftrightarrow \nu)\, d\nu \;=\; \varepsilon.
\end{aligned}$$

For the second range of integration, consider the integral

$$\begin{aligned}\left|\int_{f+\Delta f}^{\infty} [S_X(f) \Leftrightarrow S_X(\nu)]\, \delta_T(f \Leftrightarrow \nu)\, d\nu\right| & \\ \leq & \int_{f+\Delta f}^{\infty} |S_X(f) \Leftrightarrow S_X(\nu)|\, \delta_T(f \Leftrightarrow \nu)\, d\nu \\ \leq & \int_{f+\Delta f}^{\infty} (|S_X(f)| + |S_X(\nu)|)\, \delta_T(f \Leftrightarrow \nu)\, d\nu \\ = & |S_X(f)| \int_{f+\Delta f}^{\infty} \delta_T(f \Leftrightarrow \nu)\, d\nu \\ & + \int_{f+\Delta f}^{\infty} |S_X(\nu)|\, \delta_T(f \Leftrightarrow \nu)\, d\nu.\end{aligned}$$

Observe that

$$\int_{f+\Delta f}^{\infty} \delta_T(f \Leftrightarrow \nu)\, d\nu \;=\; \int_{-\infty}^{-\Delta f} \delta_T(\theta)\, d\theta,$$

which goes to zero by the second property of δ_T. Using the third property, we have

$$\begin{aligned}\int_{f+\Delta f}^{\infty} |S_X(\nu)|\, \delta_T(f \Leftrightarrow \nu)\, d\nu &= \int_{-\infty}^{-\Delta f} |S_X(f \Leftrightarrow \theta)|\, \delta_T(\theta)\, d\theta \\ &\leq \frac{1}{2T(\pi \Delta f)^2} \int_{-\infty}^{-\Delta f} |S_X(f \Leftrightarrow \theta)|\, d\theta,\end{aligned}$$

which also goes to zero as $T \to \infty$.

***Note* 3.** In applying the derivation in Note 2 to the special case $f = 0$ in (6.21) we do not need $S_X(f)$ to be continuous for all f, we only need continuity at $f = 0$.

6.9. Problems

Recall that if

$$H(f) \;=\; \int_{-\infty}^{\infty} h(t) e^{-j2\pi f t}\, dt,$$

then the **inverse Fourier transform** is

$$h(t) \;=\; \int_{-\infty}^{\infty} H(f) e^{j2\pi f t}\, df.$$

With these formulas in mind, the following table of Fourier transform pairs may be useful.

$h(t)$	$H(f)$
$I_{[-T,T]}(t)$	$2T\dfrac{\sin(2\pi Tf)}{2\pi Tf}$
$2W\dfrac{\sin(2\pi Wt)}{2\pi Wt}$	$I_{[-W,W]}(f)$
$(1 \Leftrightarrow \lvert t\rvert/T)\, I_{[-T,T]}(t)$	$T\left[\dfrac{\sin(\pi Tf)}{\pi Tf}\right]^2$
$W\left[\dfrac{\sin(\pi Wt)}{\pi Wt}\right]^2$	$(1 \Leftrightarrow \lvert f\rvert/W)\, I_{[-W,W]}(f)$
$e^{-\lambda t}u(t)$	$\dfrac{1}{\lambda + j2\pi f}$
$e^{-\lambda \lvert t\rvert}$	$\dfrac{2\lambda}{\lambda^2 + (2\pi f)^2}$
$\dfrac{\lambda}{\lambda^2 + t^2}$	$\pi\, e^{-2\pi\lambda \lvert f\rvert}$
$e^{-(t/\sigma)^2/2}$	$\sqrt{2\pi}\,\sigma\, e^{-\sigma^2(2\pi f)^2/2}$

Problems §6.1: Mean, Correlation, and Covariance

1. The Cauchy–Schwarz inequality says that for any random variables U and V,

$$\mathsf{E}[UV]^2 \;\le\; \mathsf{E}[U^2]\,\mathsf{E}[V^2]. \tag{6.24}$$

Derive this formula as follows. For any constant λ, we can always write

$$\begin{aligned} 0 &\le \mathsf{E}[\lvert U \Leftrightarrow \lambda V\rvert^2] \\ &= \mathsf{E}[U^2] \Leftrightarrow 2\lambda\mathsf{E}[UV] + \lambda^2\mathsf{E}[V^2]. \end{aligned} \tag{6.25}$$

Now set $\lambda = \mathsf{E}[UV]/\mathsf{E}[V^2]$ and rearrange the inequality. Note that since equality holds in (6.25) if and only if $U = \lambda V$, equality holds in (6.24) if and only if U is a multiple of V.

Remark. The same technique can be used to show that for complex-valued waveforms,

$$\left\lvert \int_{-\infty}^{\infty} g(\theta)\, h(\theta)^*\, d\theta \right\rvert^2 \;\le\; \int_{-\infty}^{\infty} \lvert g(\theta)\rvert^2\, d\theta \cdot \int_{-\infty}^{\infty} \lvert h(\theta)\rvert^2\, d\theta,$$

where the asterisk denotes complex conjugation, and for any complex number z, $\lvert z\rvert^2 = z \cdot z^*$.

2. Show that $R_X(t_1, t_2) = \mathsf{E}[X_{t_1} X_{t_2}]$ is a **positive semidefinite function** in the sense that for any real or complex constants $c_1, \ldots, c_n$ and any

times $t_1, \ldots, t_n$,

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_i R_X(t_i, t_k) c_k^* \;\geq\; 0.$$

Hint: Observe that

$$\mathsf{E}\left[\left|\sum_{i=1}^{n} c_i X_{t_i}\right|^2\right] \;\geq\; 0.$$

3. Let X_t for $t > 0$ be a random process with zero mean and correlation function $R_X(t_1, t_2) = \min(t_1, t_2)$. If X_t is Gaussian for each t, write down the density of X_t.

Problems §6.2: Wide-Sense Stationary Processes

4. Find the correlation function corresponding to each of the following power spectral densities. (a) $\delta(f)$. (b) $\delta(f \Leftrightarrow f_0) + \delta(f + f_0)$. (c) $e^{-f^2/2}$. (d) $e^{-|f|}$.

5. Let X_t be a WSS random process with power spectral density $S_X(f) = I_{[-W,W]}(f)$. Find $\mathsf{E}[X_t^2]$.

6. Explain why each of the following frequency functions cannot be a power spectral density. (a) $e^{-f}\,u(f)$, where u is the unit step function. (b) $e^{-f^2}\cos(f)$. (c) $(1 \Leftrightarrow f^2)/(1 + f^4)$. (d) $1/(1 + jf^2)$.

7. For each of the following functions, determine whether or not it is a valid correlation function.
(a) $\sin(\tau)$. (b) $\cos(\tau)$. (c) $e^{-\tau^2/2}$. (d) $e^{-|\tau|}$. (e) $\tau^2 e^{-|\tau|}$. (f) $I_{[-T,T]}(\tau)$.

8. Let $R_0(\tau)$ be a correlation function, and put $R(\tau) := R_0(\tau)\cos(2\pi f_0 \tau)$ for some $f_0 > 0$. Determine whether or not $R(\tau)$ is a valid correlation function.

*9. Let $S(f)$ be a real-valued, even, nonnegative function, and put

$$R(\tau) \;:=\; \int_{-\infty}^{\infty} S(f) e^{j\,2\pi f \tau}\, d\tau.$$

Show that R satisfies properties (i) and (ii) that characterize a correlation function.

*10. Let $R_0(\tau)$ be a real-valued, even function, but not necessarily a correlation function. Let $R(\tau)$ denote the convolution of R_0 with itself, i.e.,

$$R(\tau) \;:=\; \int_{-\infty}^{\infty} R_0(\theta)\, R_0(\tau \Leftrightarrow \theta)\, d\theta.$$

(a) Show that $R(\tau)$ is a valid correlation function. *Hint:* You will need the remark made in Problem 1.

(b) Now suppose that $R_0(\tau) = I_{[-T,T]}(\tau)$. In this case, what is $R(\tau)$, and what is its Fourier transform?

11. A discrete-time random process is WSS if $\mathsf{E}[X_n]$ does not depend on n and if the correlation $\mathsf{E}[X_{n+k}X_k]$ does not depend on k. In this case we write $R_X(n) = \mathsf{E}[X_{n+k}X_k]$. For discrete-time WSS processes, the power spectral density is defined by

$$S_X(f) \;:=\; \sum_{n=-\infty}^{\infty} R_X(n)e^{-j2\pi fn},$$

which is a periodic function of f with period one. By the formula for Fourier series coefficients

$$R_X(n) \;=\; \int_{-1/2}^{1/2} S_X(f)e^{j2\pi fn}\,df.$$

(a) Show that R_X is an even function of n.

(b) Show that S_X is a real and even function of f.

Problems §6.3: WSS Processes through LTI Systems

12. White noise with power spectral density $S_X(f) = \mathcal{N}_0/2$ is applied to a lowpass filter with transfer function

$$H(f) \;=\; \begin{cases} 1 \Leftrightarrow f^2, & |f| \le 1, \\ 0, & |f| > 1. \end{cases}$$

Find the output power of the filter.

13. White noise with power spectral density $\mathcal{N}_0/2$ is applied to a lowpass filter with transfer function $H(f) = \sin(\pi f)/(\pi f)$. Find the output noise power from the filter.

14. A WSS input signal X_t with correlation function $R_X(\tau) = e^{-\tau^2/2}$ is passed through an LTI system with transfer function $H(f) = e^{-(2\pi f)^2/2}$. Denote the system output by Y_t. Find (a) the cross power spectral density, $S_{XY}(f)$; (b) the cross-correlation, $R_{XY}(\tau)$; (c) $\mathsf{E}[X_{t_1}Y_{t_2}]$; (d) the output power spectral density, $S_Y(f)$; (e) the output auto-correlation, $R_Y(\tau)$; (f) the output power P_Y.

15. White noise with power spectral density $S_X(f) = \mathcal{N}_0/2$ is applied to a lowpass RC filter with impulse response $h(t) = \frac{1}{RC}e^{-t/(RC)}u(t)$. Find (a) the cross power spectral density, $S_{XY}(f)$; (b) the cross-correlation, $R_{XY}(\tau)$; (c) $\mathsf{E}[X_{t_1}Y_{t_2}]$; (d) the output power spectral density, $S_Y(f)$; (e) the output auto-correlation, $R_Y(\tau)$; (f) the output power P_Y.

16. White noise with power spectral density $\mathcal{N}_0/2$ is passed through a linear, time-invariant system with impulse response $h(t) = 1/(1+t^2)$. If Y_t denotes the filter output, find $\mathsf{E}[Y_{t+1/2}Y_t]$.

17. A WSS process X_t with correlation function $R_X(\tau) = 1/(1+\tau^2)$ is passed through an LTI system with impulse response $h(t) = 3\sin(\pi t)/(\pi t)$. Let Y_t denote the system output. Find the output power P_Y.

18. White noise with power spectral density $S_X(f) = \mathcal{N}_0/2$ is passed though a filter with impulse response $h(t) = I_{[-T/2,T/2]}(t)$. Find and sketch the correlation function of the filter output.

19. Let X_t be a WSS random process, and put $Y_t := \int_{t-3}^{t} X_\tau \, d\tau$. Determine whether or not Y_t is WSS.

20. Consider the system

$$Y_t \;=\; e^{-t}\int_{-\infty}^{t} e^{\theta} X_\theta \, d\theta.$$

Assume that X_t is zero mean white noise with power spectral density $S_X(f) = \mathcal{N}_0/2$. Show that X_t and Y_t are J-WSS, and find $R_{XY}(\tau)$, $S_{XY}(f)$, $S_Y(f)$, and $R_Y(\tau)$.

21. A zero-mean, WSS process X_t with correlation function $(1 \Leftrightarrow |\tau|)I_{[-1,1]}(\tau)$ is to be processed by a filter with transfer function $H(f)$ designed so that the system output Y_t has correlation function

$$R_Y(\tau) \;=\; \frac{\sin(\pi\tau)}{\pi\tau}.$$

Find a formula for, and sketch the required filter $H(f)$.

22. Let X_t be a WSS random process. Put $Y_t := \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau)X_\tau \, d\tau$, and $Z_t := \int_{-\infty}^{\infty} g(t \Leftrightarrow \theta)X_\theta \, d\theta$. Determine whether or not Y_t and Z_t are J-WSS.

23. Let X_t be a zero-mean WSS random process with power spectral density $S_X(f) = 2/[1+(2\pi f)^2]$. Put $Y_t := X_t \Leftrightarrow X_{t-1}$.

    (a) Show that X_t and Y_t are J-WSS.

    (b) Find the power spectral density $S_Y(f)$.

    (c) Find the power in Y_t.

24. Let $\{X_t\}$ be a zero-mean wide-sense stationary random process with power spectral density $S_X(f)$. Consider the process

$$Y_t \;:=\; \sum_{n=-\infty}^{\infty} h_n X_{t-n},$$

with h_n real valued.

(a) Show that $\{X_t\}$ and $\{Y_t\}$ are *jointly* wide-sense stationary.

(b) Show that $S_Y(f)$ has the form $S_Y(f) = P(f)S_X(f)$ where P is a real-valued, nonnegative, periodic function of f with period 1. Give a formula for $P(f)$.

25. *System Identification.* When white noise $\{W_t\}$ with power spectral density $S_W(f) = 3$ is applied to a certain linear time-invariant system, the output has power spectral density e^{-f^2}. Now let $\{X_t\}$ be a zero-mean, wide-sense stationary random process with power spectral density $S_X(f) = e^{f^2} I_{[-1,1]}(f)$. If $\{Y_t\}$ is the response of the system to $\{X_t\}$, find $R_Y(\tau)$ for all τ.

*26. *Extension to Complex Random Processes.* If X_t is a complex-valued random process, then its auto-correlation function is defined by $R_X(t_1,t_2) := \mathsf{E}[X_{t_1}X_{t_2}^*]$. Similarly, if Y_t is another complex-valued random process, their cross-correlation is defined by $R_{XY}(t_1,t_2) := \mathsf{E}[X_{t_1}Y_{t_2}^*]$. The concepts of WSS, J-WSS, the power spectral density, and the cross power spectral density are defined as in the real case. Now suppose that X_t is a complex WSS process and that $Y_t = \int_{-\infty}^{\infty} h(t \ominus \tau) X_\tau \, d\tau$, where the impulse response h is now possibly complex valued.

(a) Show that $R_X(\ominus\tau) = R_X(\tau)^*$.

(b) Show that $S_X(f)$ must be real valued. *Hint:* What does part (a) say about the real and imaginary parts of $R_X(\tau)$?

(c) Show that

$$\mathsf{E}[X_{t_1}Y_{t_2}^*] \;=\; \int_{-\infty}^{\infty} h(\ominus\beta)^* R_X([t_1 \ominus t_2] \ominus \beta)\, d\beta.$$

(d) Even though the above result is a little different from (6.4), show that (6.5) and (6.6) still hold for complex random processes.

27. Let X_n be a discrete-time WSS process as defined in Problem 11. Put

$$Y_n \;=\; \sum_{k=-\infty}^{\infty} h_k X_{n-k}.$$

(a) Suggest a definition of jointly WSS discrete-time processes and show that X_n and Y_n are J-WSS.

(b) Suggest a definition for the cross power spectral density of two discrete-time J-WSS processes and show that (6.5) holds if

$$H(f) \;:=\; \sum_{n=-\infty}^{\infty} h_n e^{-j2\pi fn}.$$

Also show that (6.6) holds.

(c) Write out the discrete-time analog of Example 6.9. What condition do you need to impose on W_2?

Problems §6.4: The Matched Filter

28. Determine the matched filter if the known radar pulse is $v(t) = \sin(t) I_{[0,\pi]}(t)$, and X_t is white noise with power spectral density $S_X(f) = \mathcal{N}_0/2$. For what values of t_0 is the optimal system causal?

29. Determine the matched filter if $v(t) = e^{-(t/\sqrt{2})^2/2}$, and $S_X(f) = e^{-(2\pi f)^2/2}$.

30. Derive the matched filter for a discrete-time received signal $v(n) + X_n$. *Hint:* Problems 11 and 27 may be helpful.

Problems §6.5: The Wiener Filter

31. Suppose V_t and X_t are J-WSS. Let $U_t := V_t + X_t$. Show that U_t and V_t are J-WSS.

32. Suppose $U_t = V_t + X_t$, where V_t and X_t are each zero mean and WSS. Also assume that $\mathsf{E}[V_t X_\tau] = 0$ for all t and τ. Express the Wiener filter $H(f)$ in terms of $S_V(f)$ and $S_X(f)$.

33. Using the setup of the previous problem, find the Wiener filter $H(f)$ and the corresponding impulse response $h(t)$ if $S_V(f) = 2\lambda/[\lambda^2 + (2\pi f)^2]$ and $S_X(f) = 1$.

    Remark. You may want to compare your answer with the *causal* Wiener filter found in Example 6.11.

34. Using the setup of Problem 32, suppose that the signal has correlation function $R_V(\tau) = \left(\dfrac{\sin \pi\tau}{\pi\tau}\right)^2$ and that the noise has power spectral density $S_X(f) = 1 \Leftrightarrow I_{[-1,1]}(f)$. Find the Wiener filter $H(f)$ and the corresponding impulse response $h(t)$.

35. Find the impulse response of the whitening filter $K(f)$ of Example 6.11. Is it causal?

36. Derive the Wiener filter for discrete-time J-WSS signals U_n and V_n with zero means. *Hints:* (*i*) First derive the analogous orthogonality principle. (*ii*) Problems 11 and 27 may be helpful.

Problems §6.6: *Expected Time-Average Power and the Wiener–Khinchin Theorem

37. Recall that by Problem 2, correlation functions are positive semidefinite. Use this fact to prove that the double integral in (6.19) is nonnegative, assuming that R_X is continuous. *Hint:* Since R_X is continuous, the double integral in (6.19) is a limit of Riemann sums of the form

$$\sum_i \sum_k R_X(t_i \Leftrightarrow t_k) e^{-j2\pi f(t_i - t_k)} \Delta t_i \Delta t_k.$$

38. Let Y_t be a WSS process. In each of the cases below, determine whether or not $\frac{1}{2T}\int_{-T}^{T} Y_t\,dt \to \mathsf{E}[Y_t]$ in mean square.

    (a) The covariance $C_Y(\tau) = e^{-|\tau|}$.

    (b) The covariance $C_Y(\tau) = \sin(\pi\tau)/(\pi\tau)$.

39. Let $Y_t = \cos(2\pi t + \Theta)$, where $\Theta \sim \text{uniform}[-\pi, \pi]$. As in Example 6.1, $\mathsf{E}[Y_t] = 0$. Determine whether or not

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} Y_t\,dt \;\to\; 0.$$

40. Let X_t be a zero-mean, WSS process. For fixed τ, you might expect

$$\frac{1}{2T}\int_{-T}^{T} X_{t+\tau}X_t\,dt$$

    to converge in mean square to $\mathsf{E}[X_{t+\tau}X_t] = R_X(\tau)$. Give conditions on the process X_t under which this will be true. *Hint:* Define $Y_t := X_{t+\tau}X_t$.

    Remark. When $\tau = 0$ this says that $\frac{1}{2T}\int_{-T}^{T} X_t^2\,dt$ converges in mean square to $R_X(0) = P_X = \widetilde{P}_X$.

41. Let X_t be a zero-mean, WSS process. For a fixed set $B \subset \mathbb{R}$, you might expect[††]

$$\frac{1}{2T}\int_{-T}^{T} I_B(X_t)\,dt$$

    to converge in mean square to $\mathsf{E}[I_B(X_t)] = \wp(X_t \in B)$. Give conditions on the process X_t under which this will be true. *Hint:* Define $Y_t := I_B(X_t)$.

[††]This is the fraction of time during $[-T, T]$ that $X_t \in B$. For example, we might have $B = [v_{\min}, v_{\max}]$ being the acceptable operating range of the voltage of some device. Then we would be interested in the fraction of time during $[-T, T]$ that the device is operating normally.

CHAPTER 7

Random Vectors

This chapter covers concepts that require some familiarity with linear algebra, e.g., matrix-vector multiplication, determinants, and matrix inverses. Section 7.1 defines the mean vector, covariance matrix, and characteristic function of random vectors. Section 7.2 introduces the multivariate Gaussian random vector and illustrates several of its properties. Section 7.3 discusses both linear and nonlinear minimum mean squared error estimation of random variables and random vectors. Estimators are obtained using appropriate orthogonality principles. Section 7.4 covers the Jacobian formula for finding the density of $Y = G(X)$ when the density of X is given. Section 7.5 introduces complex-valued random variables and random vectors. Emphasis is on circularly symmetric Gaussian random vectors due to their importance in the design of digital communication systems.

7.1. Mean Vector, Covariance Matrix, and Characteristic Function

If $X = [X_1, \ldots, X_n]'$ is a random vector, we put $m_i := \mathsf{E}[X_i]$, and we define R_{ij} to be the **covariance** between X_i and X_j, i.e.,

$$R_{ij} \;:=\; \mathsf{cov}(X_i, X_j) \;:=\; \mathsf{E}[(X_i \Leftrightarrow m_i)(X_j \Leftrightarrow m_j)].$$

Note that $R_{ii} = \mathsf{cov}(X_i, X_i) = \mathsf{var}(X_i)$. We call the vector $m := [m_1, \ldots, m_n]'$ the **mean vector** of X, and we call the $n \times n$ matrix $R := [R_{ij}]$ the **covariance matrix** of X. Note that since $R_{ij} = R_{ji}$, the matrix R is **symmetric**. In other words, $R' = R$. Also note that for $i \neq j$, $R_{ij} = 0$ if and only if X_i and X_j are uncorrelated. Thus, R is a diagonal matrix if and only if X_i and X_j are uncorrelated for all $i \neq j$.

If we define the expectation of a random vector $X = [X_1, \ldots, X_n]'$ to be the vector of expectations, i.e.,

$$\mathsf{E}[X] \;:=\; \begin{bmatrix} \mathsf{E}[X_1] \\ \vdots \\ \mathsf{E}[X_n] \end{bmatrix},$$

then $\mathsf{E}[X] = m$.

We define the expectation of a matrix of random variables to be the matrix of expectations. This leads to the following characterization of the covariance matrix R. To begin, observe that if we multiply an $n \times 1$ matrix (a column vector) by a $1 \times n$ matrix (a row vector), we get an $n \times n$ matrix. Hence,

$(X \Leftrightarrow m)(X \Leftrightarrow m)'$ is equal to

$$\begin{bmatrix} (X_1 \Leftrightarrow m_1)(X_1 \Leftrightarrow m_1) & \cdots & (X_1 \Leftrightarrow m_1)(X_n \Leftrightarrow m_n) \\ \vdots & & \vdots \\ (X_n \Leftrightarrow m_n)(X_1 \Leftrightarrow m_1) & \cdots & (X_n \Leftrightarrow m_n)(X_n \Leftrightarrow m_n) \end{bmatrix}.$$

The expectation of the ijth entry is $\mathsf{cov}(X_i, X_j) = R_{ij}$. Thus,

$$R \;=\; \mathsf{E}[(X \Leftrightarrow m)(X \Leftrightarrow m)'] \;=:\; \mathsf{cov}(X).$$

Note the distinction between the covariance of a pair of random variables, which is a scalar, and the covariance of a column vector, which is a matrix.

***Example* 7.1.** Write out the covariance matrix of the three-dimensional random vector $U := [X, Y, Z]'$ if U has zero mean.

Solution. The covariance matrix of U is

$$\mathsf{cov}(U) \;=\; \mathsf{E}[UU'] \;=\; \begin{bmatrix} \mathsf{E}[X^2] & \mathsf{E}[XY] & \mathsf{E}[XZ] \\ \mathsf{E}[YX] & \mathsf{E}[Y^2] & \mathsf{E}[YZ] \\ \mathsf{E}[ZX] & \mathsf{E}[ZY] & \mathsf{E}[Z^2] \end{bmatrix}.$$

***Example* 7.2.** Let $X = [X_1, \ldots, X_n]'$ be a random vector with covariance matrix R, and let $c = [c_1, \ldots, c_n]'$ be a given constant vector. Define the scalar

$$Z \;:=\; c'(X \Leftrightarrow m) \;=\; \sum_{i=1}^{n} c_i (X_i \Leftrightarrow m_i),$$

Show that

$$\mathsf{var}(Z) \;=\; c'Rc.$$

Solution. First note that since $\mathsf{E}[X] = m$, Z has zero mean. Hence, $\mathsf{var}(Z) = \mathsf{E}[Z^2]$. Write

$$\begin{aligned} \mathsf{E}[Z^2] &= \mathsf{E}\left[\left(\sum_{i=1}^{n} c_i (X_i \Leftrightarrow m_i)\right)\left(\sum_{j=1}^{n} c_j (X_j \Leftrightarrow m_j)\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \mathsf{E}[(X_i \Leftrightarrow m_i)(X_j \Leftrightarrow m_j)] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j R_{ij} \\ &= \sum_{i=1}^{n} c_i \left(\sum_{j=1}^{n} R_{ij} c_j\right). \end{aligned}$$

The inner sum is the ith component of the column vector Rc. Hence, $\mathsf{E}[Z^2] = c'(Rc) = c'Rc$.

The preceding example shows that a covariance matrix must satisfy

$$c'Rc \;=\; \mathsf{E}[Z^2] \;\geq\; 0$$

for all vectors $c = [c_1, \ldots, c_n]'$. A symmetric matrix R with the property $c'Rc \geq 0$ for all vectors c is said to be **positive semidefinite**. If $c'Rc > 0$ for all nonzero c, then R is called **positive definite**.

Recall that the **norm** of a column vector x is defined by $\|x\| := (x'x)^{1/2}$. It is shown in Problem 5 that

$$\mathsf{E}[\|X \Leftrightarrow \mathsf{E}[X]\|^2] \;=\; \mathrm{tr}(R) \;=\; \sum_{i=1}^{n} \mathsf{var}(X_i),$$

where the **trace** of an $n \times n$ matrix R is defined by

$$\mathrm{tr}(R) \;:=\; \sum_{i=1}^{n} R_{ii}.$$

The **joint characteristic function** of $X = [X_1, \ldots, X_n]'$ is defined by

$$\varphi_X(\nu) \;:=\; \mathsf{E}[e^{j\nu'X}] \;=\; \mathsf{E}[e^{j(\nu_1 X_1 + \cdots + \nu_n X_n)}],$$

where $\nu = [\nu_1, \ldots, \nu_n]'$.

Note that when X has a joint density, $\varphi_X(\nu) = \mathsf{E}[e^{j\nu'X}]$ is just the n-dimensional Fourier transform,

$$\varphi_X(\nu) \;=\; \int_{\mathbb{R}^n} e^{j\nu'x} f_X(x)\, dx, \tag{7.1}$$

and the joint density can be recovered using the multivariate inverse Fourier transform:

$$f_X(x) \;=\; \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-j\nu'x} \varphi_X(\nu)\, d\nu.$$

Whether X has a joint density or not, the joint characteristic function can be used to obtain its various moments.

***Example* 7.3.** The components of the mean vector and covariance matrix can be obtained from the characteristic function as follows. Write

$$\frac{\partial}{\partial \nu_k} \mathsf{E}[e^{j\nu'X}] \;=\; \mathsf{E}[e^{j\nu'X} j X_k],$$

and

$$\frac{\partial^2}{\partial \nu_\ell \partial \nu_k} \mathsf{E}[e^{j\nu'X}] \;=\; \mathsf{E}[e^{j\nu'X} (jX_\ell)(jX_k)].$$

Then

$$\frac{\partial}{\partial \nu_k} \mathsf{E}[e^{j\nu'X}]\Big|_{\nu=0} \;=\; j\mathsf{E}[X_k],$$

and

$$\frac{\partial^2}{\partial \nu_\ell \partial \nu_k} \mathsf{E}[e^{j\nu' X}]\Big|_{\nu=0} = \mathsf{E}[X_\ell X_k].$$

Higher-order moments can be obtained in a similar fashion.

If the components of $X = [X_1, \ldots, X_n]'$ are independent, then

$$\begin{aligned}
\varphi_X(\nu) &= \mathsf{E}[e^{j\nu' X}] \\
&= \mathsf{E}[e^{j(\nu_1 X_1 + \cdots + \nu_n X_n)}] \\
&= \mathsf{E}\left[\prod_{k=1}^{n} e^{j\nu_k X_k}\right] \\
&= \prod_{k=1}^{n} \mathsf{E}[e^{j\nu_k X_k}] \\
&= \prod_{k=1}^{n} \varphi_{X_k}(\nu_k).
\end{aligned}$$

In other words, the joint characteristic function is the product of the marginal characteristic functions.

7.2. The Multivariate Gaussian

A random vector $X = [X_1, \ldots, X_n]'$ is said to be Gaussian or normal if every linear combination of the components of X, e.g.,

$$\sum_{i=1}^{n} c_i X_i, \tag{7.2}$$

is a scalar Gaussian random variable.

***Example* 7.4.** If X is a Gaussian random vector, then the numerical average of its components,

$$\frac{1}{n} \sum_{i=1}^{n} X_i,$$

is a scalar Gaussian random variable.

An easy consequence of our definition of Gaussian random vector is that any subvector is also Gaussian. To see this, suppose $X = [X_1, \ldots, X_n]'$ is a Gaussian random vector. Then every linear combination of the components of the subvector $[X_1, X_3, X_5]'$ is of the form (7.2) if we take $c_i = 0$ for i not equal to $1, 3, 5$.

Sometimes it is more convenient to express linear combinations as the product of a row vector times the column vector X. For example, if we put $c = [c_1, \ldots, c_n]'$, then

$$\sum_{i=1}^{n} c_i X_i \;=\; c'X.$$

Now suppose that $Y = AX$ for some $p \times n$ matrix A. Letting $c = [c_1, \ldots, c_p]'$, every linear combination of the p components of Y has the form

$$\sum_{i=1}^{p} c_i Y_i \;=\; c'Y \;=\; c'(AX) \;=\; (A'c)'X,$$

which is a linear combination of the components of X, and therefore normal.

We can even add a constant vector. If $Y = AX + b$, where A is again $p \times n$, and b is $p \times 1$, then

$$c'Y \;=\; c'(AX + b) \;=\; (A'c)'X + c'b.$$

Adding the constant $c'b$ to the normal random variable $(A'c)'X$ results in another normal random variable (with a different mean).

In summary, *if X is a Gaussian random vector, then so is $AX + b$ for any $p \times n$ matrix A and any p-vector b.*

The Characteristic Function of a Gaussian Random Vector

We now find the joint characteristic function of the Gaussian random vector X, $\varphi_X(\nu) = \mathsf{E}[e^{j\nu'X}]$. Since we are assuming that X is a normal vector, the quantity $Y := \nu'X$ is a scalar Gaussian random variable. Hence,

$$\varphi_X(\nu) \;=\; \mathsf{E}[e^{j\nu'X}] \;=\; \mathsf{E}[e^{jY}] \;=\; \mathsf{E}[e^{j\eta Y}]\big|_{\eta=1} \;=\; \varphi_Y(\eta)\big|_{\eta=1}.$$

In other words, all we have to do is find the characteristic function of Y and evaluate it at $\eta = 1$. Since Y is normal, we know its characteristic function is (recall Example 3.13)

$$\varphi_Y(\eta) \;:=\; \mathsf{E}[e^{j\eta Y}] \;=\; e^{j\eta\mu - \eta^2\sigma^2/2},$$

where $\mu := \mathsf{E}[Y]$, and $\sigma^2 := \mathsf{var}(Y)$. Assuming X has mean vector m, $\mu = \mathsf{E}[\nu'X] = \nu'm$. Suppose X has covariance matrix R. Next, write $\mathsf{var}(Y) = \mathsf{E}[(Y \Leftrightarrow \mu)^2]$. Since

$$Y \Leftrightarrow \mu \;=\; \nu'(X \Leftrightarrow m),$$

Example 7.2 tells us that $\mathsf{var}(Y) \;=\; \nu'R\nu$. It now follows that

$$\varphi_X(\nu) \;=\; e^{j\nu'm - \nu'R\nu/2}.$$

If X is a Gaussian random vector with mean vector m and covariance matrix R, we write $X \sim N(m, R)$.

For Gaussian Random Vectors, Uncorrelated Implies Independent

If the components of a random vector are uncorrelated, then the covariance matrix is diagonal. In general, this is not enough to prove that the components of the random vector are independent. However, if X is a Gaussian random vector, then the components are independent. To see this, suppose that X is Gaussian with uncorrelated components. Then R is diagonal, say

$$R = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix},$$

where $\sigma_i^2 = R_{ii} = \mathsf{var}(X_i)$. The diagonal form of R implies that

$$\nu' R \nu = \sum_{i=1}^{n} \sigma_i^2 \nu_i^2,$$

and so

$$\varphi_X(\nu) = e^{j\nu' m - \nu' R \nu/2} = \prod_{i=1}^{n} e^{j\nu_i m_i - \sigma_i^2 \nu_i^2/2}.$$

In other words,

$$\varphi_X(\nu) = \prod_{i=1}^{n} \varphi_{X_i}(\nu_i),$$

where $\varphi_{X_i}(\nu_i)$ is the characteristic function of the $N(m_i, \sigma_i^2)$ density. Multivariate inverse Fourier transformation then yields

$$f_X(x) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

where $f_{X_i} \sim N(m_i, \sigma_i^2)$. This establishes the independence of the X_i.

***Example* 7.5.** Let $X_1, \ldots, X_n$ be i.i.d. $N(m, \sigma^2)$ random variables. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the average of the X_i. Furthermore, for $j = 1, \ldots, n$, put $Y_j := X_j \Leftrightarrow \overline{X}$. Show that $\overline{X}$ and $Y := [Y_1, \ldots, Y_n]'$ are jointly normal and independent.

Solution. Let $X := [X_1, \ldots, X_n]'$, and put $a := [\frac{1}{n} \cdots \frac{1}{n}]$. Then $\overline{X} = aX$. Next, observe that

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \Leftrightarrow \begin{bmatrix} \overline{X} \\ \vdots \\ \overline{X} \end{bmatrix}.$$

Let M denote the $n \times n$ matrix with each row equal to a; i.e., $M_{ij} = 1/n$ for all i, j. Then $Y = X \Leftrightarrow MX = (I \Leftrightarrow M)X$, and Y is a jointly normal random vector. Next consider the vector

$$Z := \begin{bmatrix} \overline{X} \\ Y \end{bmatrix} = \begin{bmatrix} a \\ I \Leftrightarrow M \end{bmatrix} X.$$

Since Z is a linear transformation of the Gaussian random vector X, Z is also a Gaussian random vector. Furthermore, its covariance matrix has the block-diagonal form (see Problem 11)

$$\begin{bmatrix} \mathsf{var}(\overline{X}) & 0 \\ 0 & \mathsf{E}[YY'] \end{bmatrix}.$$

This implies, by Problem 12, that $\overline{X}$ and Y are independent.

The Density Function of a Gaussian Random Vector

We now derive the general multivariate Gaussian density function under the assumption that R is positive definite. Using the multivariate Fourier inversion formula,

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-jx'\nu} \varphi_X(\nu)\, d\nu \\ &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-jx'\nu} e^{j\nu' m - \nu' R\nu/2}\, d\nu \\ &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-j(x-m)'\nu} e^{-\nu' R\nu/2}\, d\nu. \end{aligned}$$

Now make the multivariate change of variable $\zeta = R^{1/2}\nu$, $d\zeta = \det R^{1/2}\, d\nu$. The existence of $R^{1/2}$ and the fact that $\det(R^{1/2}) = \sqrt{\det R} > 0$ are shown in the Notes.[1] We have

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-j(x-m)'R^{-1/2}\zeta} e^{-\zeta'\zeta/2} \frac{d\zeta}{\det R^{1/2}} \\ &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-j\{R^{-1/2}(x-m)\}'\zeta} e^{-\zeta'\zeta/2} \frac{d\zeta}{\sqrt{\det R}}. \end{aligned}$$

Put $t = R^{-1/2}(x \Leftrightarrow m)$. Then

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-jt'\zeta} e^{-\zeta'\zeta/2} \frac{d\zeta}{\sqrt{\det R}} \\ &= \frac{1}{\sqrt{\det R}} \prod_{i=1}^{n} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jt_i\zeta_i} e^{-\zeta_i^2/2}\, d\zeta_i \right). \end{aligned}$$

Observe that $e^{-\zeta_i^2/2}$ is the characteristic function of a scalar $N(0,1)$ random variable. Hence,

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{\det R}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-t_i^2/2} \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det R}} \exp\left[\textstyle\frac{1}{2}\sum_{i=1}^{n} t_i^2\right] \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det R}} \exp[t't/2] \\ &= \frac{\exp[\frac{1}{2}(x-m)'R^{-1}(x-m)]}{(2\pi)^{n/2}\sqrt{\det R}}. \end{aligned} \tag{7.3}$$

With the norm notation, $t't = \|t\|^2 = \|R^{-1/2}(x-m)\|^2$, and so

$$f_X(x) = \frac{\exp[\|R^{-1/2}(x-m)\|^2/2]}{(2\pi)^{n/2}\sqrt{\det R}}$$

as well.

7.3. Estimation of Random Vectors

Consider a pair of random vectors X and Y, where X is not observed, but Y is observed. We wish to estimate X based on our knowledge of Y. By an **estimator** of X based on Y, we mean a function $g(y)$ such that $\hat{X} := g(Y)$ is our **estimate** or "guess" of the value of X. What is the best function g to use? What do we mean by best? In this section, we define g to be best if it minimizes the **mean squared error** (MSE) $\mathsf{E}[\|X - g(Y)\|^2]$ for all functions g in some class of functions. The optimal function g is called the **minimum mean squared error** (MMSE) estimator. In the next subsection, we restrict attention to the class of **linear estimators** (actually affine). Later we allow g to be arbitrary.

Linear Minimum Mean Squared Error Estimation

We now restrict attention to estimators g of the form $g(y) = Ay + b$, where A is a matrix and b is a column vector. Such a function of y is said to be affine. If $b = 0$, then g is linear. It is common to say g is linear even if $b \neq 0$ since this only is a slight abuse of terminology, and the meaning is understood. We shall follow this convention. To find the best linear estimator is simply to find the matrix A and the column vector b that minimize the MSE, which for linear estimators has the form

$$\mathsf{E}\left[\left\| X - (AY + b)\right\|^2\right].$$

Letting $m_X = \mathsf{E}[X]$ and $m_Y = \mathsf{E}[Y]$, the MSE is equal to

$$\mathsf{E}\left[\left\| \left\{(X - m_X) - A(Y - m_Y)\right\} + \left\{m_X - Am_Y - b\right\}\right\|^2\right].$$

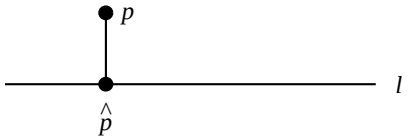

Figure 7.1. Geometric interpretation of the orthogonality principle.

Since the left-hand quantity in braces is zero mean, and since the right-hand quantity in braces is a constant (nonrandom), the MSE simplifies to

$$\mathsf{E}\big[\big\|(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)\big\|^2\big] + \|m_X \Leftrightarrow Am_Y \Leftrightarrow b\|^2.$$

No matter what matrix A is used, the optimal choice of b is

$$b \;=\; m_X \Leftrightarrow Am_Y\,,$$

and the estimate is $g(Y) = AY + b = A(Y \Leftrightarrow m_Y) + m_X$. The estimate is truly linear in Y if and only if $Am_Y = m_X$.

We now turn to the problem of minimizing

$$\mathsf{E}\big[\big\|(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)\big\|^2\big].$$

The matrix A is optimal if and only if for all matrices B,

$$\mathsf{E}\big[\big\|(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)\big\|^2\big] \le \mathsf{E}\big[\big\|(X \Leftrightarrow m_X) \Leftrightarrow B(Y \Leftrightarrow m_Y)\big\|^2\big]. \tag{7.4}$$

The following condition is equivalent, and easier to use. This equivalence is known as the **orthogonality principle**. It says that (7.4) holds for all B if and only if

$$\mathsf{E}\big[\big\{B(Y \Leftrightarrow m_Y)\big\}'\big\{(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)\big\}\big] \;=\; 0, \quad \text{for all } B. \tag{7.5}$$

Below we prove that (7.5) implies (7.4). The converse is also true, but we shall not use it in this book.

We first explain the terminology and show geometrically why it is true. Recall that given a straight line l and a point p not on the line, the shortest path between p and the line is obtained by dropping a perpendicular segment from the point to the line as shown in Figure 7.1. The point $\hat{p}$ on the line where the segment touches is closer to p than any other point on the line. Notice also that the vertical segment, $p \Leftrightarrow \hat{p}$, is orthogonal to the line l. In our situation, the role of p is played by the random variable $X \Leftrightarrow m_X$, the role of $\hat{p}$ is played by the random variable $A(Y \Leftrightarrow m_Y)$, and the role of the line is played by the set of all random variables of the form $B(Y \Leftrightarrow m_Y)$ as B runs over all matrices. Since the inner product between two random vectors U and V can be defined as $\mathsf{E}[V'U]$, (7.5) says that

$$(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)$$

is orthogonal to all $B(Y \Leftrightarrow m_Y)$.

To use (7.5), first note that since it is a scalar equation, the left-hand side is equal to its trace. Bringing the trace inside the expectation and using the fact that $\mathrm{tr}(\alpha\beta) = \mathrm{tr}(\beta\alpha)$, we see that the left-hand side of (7.5) is equal to

$$\mathsf{E}\big[\mathrm{tr}\big(\big\{(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)\big\}(Y \Leftrightarrow m_Y)'B'\big)\big].$$

Taking the trace back out of the expectation shows that (7.5) is equivalent to

$$\mathrm{tr}([R_{XY} \Leftrightarrow AR_Y]B') \;=\; 0, \quad \text{for all } B, \tag{7.6}$$

where $R_Y = \mathsf{cov}(Y)$ is the covariance matrix of Y, and

$$R_{XY} \;=\; \mathsf{E}[(X \Leftrightarrow m_X)(Y \Leftrightarrow m_Y)'] \;=:\; \mathsf{cov}(X,Y).$$

(This is our third usage of cov. We have already seen the covariance of a pair of random variables and of a random vector. This is the third type of covariance, that of a pair of random vectors. When X and Y are both of dimension one, this reduces to the first usage of cov.) By Problem 5, it follows that (7.6) holds if and only if A solves the equation

$$AR_Y \;=\; R_{XY}.$$

If R_Y is invertible, the unique solution of this equation is

$$A \;=\; R_{XY}R_Y^{-1}.$$

In this case, the complete estimate of X is

$$R_{XY}R_Y^{-1}(Y \Leftrightarrow m_Y) + m_X.$$

Even if R_Y is not invertible, $AR_Y = R_{XY}$ always has a solution, as shown in Problem 16.

***Example* 7.6** (Signal in Additive Noise). Let X denote a random signal of zero mean and known covariance matrix R_X. Suppose that in order to estimate X, all we have available is the noisy measurement

$$Y \;=\; X + W,$$

where W is a noise vector with zero mean and known covariance matrix R_W. Further assume that the covariance between the signal and noise, R_{XW}, is zero. Find the linear MMSE estimate of X based on Y.

Solution. Since $\mathsf{E}[Y] = \mathsf{E}[X + W] = 0$, $m_Y = m_X = 0$. Next,

$$\begin{aligned} R_{XY} &= \mathsf{E}[(X \Leftrightarrow m_X)(Y \Leftrightarrow m_Y)'] \\ &= \mathsf{E}[X(X+W)'] \\ &= R_X. \end{aligned}$$

Similarly,

$$\begin{aligned} R_Y &= \mathsf{E}[(Y \Leftrightarrow m_Y)(Y \Leftrightarrow m_Y)'] \\ &= \mathsf{E}[(X+W)(X+W)'] \\ &= R_X + R_W. \end{aligned}$$

It follows that

$$\hat{X} = R_X(R_X + R_W)^{-1}Y.$$

We now show that (7.5) implies (7.4). To simplify the notation, we assume zero means.

$$\begin{aligned} \mathsf{E}[\|X \Leftrightarrow BY\|^2] &= \mathsf{E}[\|(X \Leftrightarrow AY) + (AY \Leftrightarrow BY)\|^2] \\ &= \mathsf{E}[\|(X \Leftrightarrow AY) + (A \Leftrightarrow B)Y)\|^2] \\ &= \mathsf{E}[\|X \Leftrightarrow AY\|^2] + \mathsf{E}[\|(A \Leftrightarrow B)Y\|^2], \end{aligned}$$

where the cross terms $2\mathsf{E}[\{(A \Leftrightarrow B)Y\}'(X \Leftrightarrow AY)]$ vanish by (7.5). If we drop the right-hand term in the above display, we obtain

$$\mathsf{E}[\|X \Leftrightarrow BY\|^2] \;\geq\; \mathsf{E}[\|X \Leftrightarrow AY\|^2].$$

Minimum Mean Squared Error Estimation

We begin with an observed vector Y and a scalar X to be estimated. (The generalization to vector-valued X is left to the problems.) We seek a real-valued function $g(y)$ with the property that*

$$\mathsf{E}[|X \Leftrightarrow g(Y)|^2] \;\leq\; \mathsf{E}[|X \Leftrightarrow h(Y)|^2], \quad \text{for all } h. \tag{7.7}$$

Here we are not requiring g and/or h to be linear. We again have an **orthogonality principle** (see Problem 17) that says that the above condition is equivalent to

$$\mathsf{E}[\{X \Leftrightarrow g(Y)\}h(Y)] \;=\; 0, \quad \text{for all } h. \tag{7.8}$$

We claim that $g(y) = \mathsf{E}[X|Y=y]$ is the optimal function g if X and Y are jointly continuous. In this case, we can use the law of total probability to write

$$\mathsf{E}[\{X \Leftrightarrow g(Y)\}h(Y)] = \int \mathsf{E}\big[\{X \Leftrightarrow g(Y)\}h(Y)\big|Y=y\big] f_Y(y)\,dy$$

Applying the substitution law to the conditional expectation yields

$$\begin{aligned} \mathsf{E}\big[\{X \Leftrightarrow g(Y)\}h(Y)\big|Y=y\big] &= \mathsf{E}\big[\{X \Leftrightarrow g(y)\}h(y)\big|Y=y\big] \\ &= \mathsf{E}[X|Y=y]h(y) \Leftrightarrow g(y)h(y) \\ &= g(y)h(y) \Leftrightarrow g(y)h(y) \\ &= 0. \end{aligned}$$

*In order that the expectations in (7.7) be finite, we assume $\mathsf{E}[X^2] < \infty$, and we restrict g and h to be such that $\mathsf{E}[g(Y)^2] < \infty$ and $\mathsf{E}[h(Y)^2] < \infty$.

It is shown in Problem 18 that the solution g of (7.8) is unique. We can use this fact to find conditional expectations if X and Y are jointly normal. For simplicity, assume zero means. We claim that $g(y) = R_{XY}R_Y^{-1}y$ solves (7.8). We first observe that

$$\begin{bmatrix} X \Leftrightarrow R_{XY}R_Y^{-1}Y \\ Y \end{bmatrix} = \begin{bmatrix} I & \Leftrightarrow R_{XY}R_Y^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

is a linear transformation of $[X, Y']'$ and so the left-hand side is a Gaussian random vector whose top and bottom entries are easily seen to be uncorrelated:

$$\begin{aligned} \mathsf{E}[(X \Leftrightarrow R_{XY}R_Y^{-1}Y)Y'] &= R_{XY} \Leftrightarrow R_{XY}R_Y^{-1}R_Y \\ &= 0. \end{aligned}$$

Being jointly Gaussian and uncorrelated, they are independent. Hence, for any function $h(y)$,

$$\begin{aligned} \mathsf{E}[(X \Leftrightarrow R_{XY}R_Y^{-1}Y)h(Y)] &= \mathsf{E}[X \Leftrightarrow R_{XY}R_Y^{-1}Y]\,\mathsf{E}[h(Y)] \\ &= 0\,\mathsf{E}[h(Y)]. \end{aligned}$$

We have now learned two things about jointly normal random variables. First, we have learned how to find conditional expectations. Second, in looking for the best estimator function g, and not requiring g to be linear, we found that the best g is linear if X and Y are jointly normal.

7.4. Transformations of Random Vectors

If $G(x)$ is a vector-valued function of $x \in \mathbb{R}^n$, and X is an $\mathbb{R}^n$-valued random vector, we can define a new random vector by $Y = G(X)$. If X has joint density f_X, and G is a suitable invertible mapping, then we can find a relatively explicit formula for the joint density of Y. Suppose that the entries of the vector equation $y = G(x)$ are given by

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} g_1(x_1, \ldots, x_n) \\ \vdots \\ g_n(x_1, \ldots, x_n) \end{bmatrix}.$$

If G is invertible, we can apply G^{-1} to both sides of $y = G(x)$ to obtain $G^{-1}(y) = x$. Using the notation $H(y) := G^{-1}(y)$, we can write the entries of the vector equation $x = H(y)$ as

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} h_1(y_1, \ldots, y_n) \\ \vdots \\ h_n(y_1, \ldots, y_n) \end{bmatrix}.$$

Assuming that H is continuous and has continuous partial derivatives, let

$$H'(y) := \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial h_i}{\partial y_1} & \cdots & \frac{\partial h_i}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{bmatrix}.$$

To compute $\wp(Y \in C) = \wp(G(X) \in C)$, it is convenient to put $B = \{x : G(x) \in C\}$ so that

$$\begin{aligned} \wp(Y \in C) &= \wp(G(X) \in C) \\ &= \wp(X \in B) \\ &= \int_{\mathbb{R}^n} I_B(x)\, f_X(x)\, dx. \end{aligned}$$

Now apply the multivariate change of variable $x = H(y)$. Keeping in mind that $dx = |\det H'(y)|\, dy$,

$$\wp(Y \in C) = \int_{\mathbb{R}^n} I_B(H(y))\, f_X(H(y))\, |\det H'(y)|\, dy.$$

Observe that $I_B(H(y)) = 1$ if and only if $H(y) \in B$, which happens if and only if $G(H(y)) \in C$. However, since $H = G^{-1}$, $G(H(y)) = y$, and we see that $I_B(H(y)) = I_C(y)$. Thus,

$$\wp(Y \in C) = \int_C f_X(H(y))\, |\det H'(y)|\, dy.$$

It follows that Y has density

$$f_Y(y) = f_X(H(y))\, |\det H'(y)|.$$

Since $\det H'(y)$ is called the **Jacobian** of H, the preceding equations are sometimes called **Jacobian formulas**.

***Example* 7.7.** Let X and Y be independent univariate $N(0,1)$ random variables. Let $R = \sqrt{X^2 + Y^2}$ and $\Theta = \tan^{-1}(Y/X)$. Find the joint density of R and Θ.

Solution. The transformation G is given by

$$\begin{aligned} r &= \sqrt{x^2 + y^2}, \\ \theta &= \tan^{-1}(y/x). \end{aligned}$$

The first thing we must do is find the inverse transformation H. To begin, observe that $x \tan\theta = y$. Also, $r^2 = x^2 + y^2$. Write

$$x^2 \tan^2\theta = y^2 = r^2 \Leftrightarrow x^2.$$

Then $x^2(1+\tan^2\theta) = r^2$. Since $1+\tan^2\theta = \sec^2\theta$,

$$x^2 \;=\; r^2\cos^2\theta.$$

It then follows that $y^2 = r^2 \Leftrightarrow x^2 = r^2(1 \Leftrightarrow \cos^2\theta) = r^2\sin^2\theta$. Hence, the inverse transformation H is given by

$$\begin{aligned} x &= r\cos\theta, \\ y &= r\sin\theta. \end{aligned}$$

The matrix $H'(r,\theta)$ is given by

$$H'(r,\theta) \;=\; \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & \Leftrightarrow r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix},$$

and $\det H'(r,\theta) = r\cos^2\theta + r\sin^2\theta = r$. Then

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= f_{XY}(x,y)\Big|_{\substack{x=r\cos\theta \\ y=r\sin\theta}} \cdot |\det H'(r,\theta)| \\ &= f_{XY}(r\cos\theta, r\sin\theta)\, r. \end{aligned}$$

Now, since X and Y are independent $N(0,1)$, $f_{XY}(x,y) = f_X(x)\,f_Y(y) = e^{-(x^2+y^2)/2}/(2\pi)$, and

$$f_{R,\Theta}(r,\theta) \;=\; re^{-r^2/2}\cdot\frac{1}{2\pi}, \quad r \geq 0, \Leftrightarrow\pi < \theta \leq \pi.$$

Thus, R and Θ are independent, with R having a Rayleigh density and Θ having a uniform$(\Leftrightarrow\pi,\pi]$ density.

7.5. Complex Random Variables and Vectors

A **complex random variable** is a pair of real random variables, say X and Y, written in the form $Z = X + jY$, where j denotes the square root of $\Leftrightarrow 1$. The advantage of the complex notation is that it becomes easy to write down certain functions of (X,Y). For example, it is easier to talk about

$$Z^2 \;=\; (X+jY)(X+jY) \;=\; (X^2 \Leftrightarrow Y^2) + j(2XY)$$

than the vector-valued mapping

$$g(X,Y) \;=\; \begin{bmatrix} X^2 \Leftrightarrow Y^2 \\ 2XY \end{bmatrix}.$$

Recall that the absolute value of a complex number $z = x + jy$ is

$$|z| \;:=\; \sqrt{x^2+y^2}.$$

The **complex conjugate** of z is

$$z^* \;:=\; x \Leftrightarrow jy,$$

and so

$$zz^* \;=\; (x+jy)(x \Leftrightarrow jy) \;=\; x^2+y^2 \;=\; |z|^2.$$

The expected value of Z is simply

$$\mathsf{E}[Z] \;:=\; \mathsf{E}[X] + j\mathsf{E}[Y].$$

The covariance of Z is

$$\mathsf{cov}(Z) \;:=\; \mathsf{E}[(Z \Leftrightarrow \mathsf{E}[Z])(Z \Leftrightarrow \mathsf{E}[Z])^*] \;=\; \mathsf{E}\big[\big|Z \Leftrightarrow \mathsf{E}[Z]\big|^2\big].$$

Note that $\mathsf{cov}(Z) = \mathsf{var}(X) + \mathsf{var}(Y)$, while

$$\mathsf{E}[(Z \Leftrightarrow \mathsf{E}[Z])^2] \;=\; [\mathsf{var}(X) \Leftrightarrow \mathsf{var}(Y)] + j[2\,\mathsf{cov}(X,Y)],$$

which is zero if and only if X and Y are uncorrelated and have the same variance.

If X and Y are jointly continuous real random variables, then we say that $Z = X + jY$ is a continuous complex random variable with density

$$f_Z(z) \;=\; f_Z(x+jy) \;:=\; f_{XY}(x,y).$$

Sometimes the formula for $f_{XY}(x,y)$ is more easily expressed in terms of the complex variable z. For example, if X and Y are independent $N(0,1/2)$, then

$$f_{XY}(x,y) \;=\; \frac{e^{-x^2}}{\sqrt{2\pi}\sqrt{1/2}} \cdot \frac{e^{-y^2}}{\sqrt{2\pi}\sqrt{1/2}} \;=\; \frac{e^{-|z|^2}}{\pi}$$

Note that $\mathsf{E}[Z] = 0$ and $\mathsf{cov}(Z) = 1$. Also, the density is circularly symmetric since $|z|^2 = x^2 + y^2$ depends only on the distance from the origin of the point $(x,y) \in \mathbb{R}^2$.

A **complex random vector** of dimension n, say

$$Z \;=\; [Z_1, \ldots, Z_n]',$$

is a vector whose ith component is a complex random variable $Z_i = X_i + jY_i$, where X_i and Y_j are real random variables. If we put

$$X \;:=\; [X_1, \ldots, X_n]' \quad \text{and} \quad Y \;:=\; [Y_1, \ldots, Y_n]',$$

then $Z = X + jY$, and the mean vector of Z is $\mathsf{E}[Z] = \mathsf{E}[X] + j\mathsf{E}[Y]$. The covariance matrix of Z is

$$C \;:=\; \mathsf{E}[(Z \Leftrightarrow \mathsf{E}[Z])(Z \Leftrightarrow \mathsf{E}[Z])^H],$$

where the superscript ^{H} denotes the complex conjugate transpose. In other words, the $i\,k$ entry of C is

$$C_{ik} \;=\; \mathsf{E}[(Z_i \Leftrightarrow \mathsf{E}[Z_i])(Z_k \Leftrightarrow \mathsf{E}[Z_k])^*] \;=:\; \mathsf{cov}(Z_i, Z_k).$$

It is also easy to show that

$$C \;=\; (R_X + R_Y) + j(R_{YX} \Leftrightarrow R_{XY}). \tag{7.9}$$

For joint distribution purposes, we identify the n-dimensional complex vector Z with the $2n$-dimensional real random vector

$$[X_1, \ldots, X_n, Y_1, \ldots, Y_n]'. \tag{7.10}$$

If this $2n$-dimensional real random vector has a joint density f_{XY}, then we write

$$f_Z(z) \;:=\; f_{XY}(x_1, \ldots, x_n, y_1, \ldots, y_n).$$

Sometimes the formula for the right-hand side can be written simply in terms of the complex vector z.

Complex Gaussian Random Vectors

An n-dimensional complex random vector $Z = X + jY$ is said to be Gaussian if the $2n$-dimensional real random vector in (7.10) is jointly Gaussian; i.e., its characteristic function $\varphi_{XY}(\nu, \theta) = \mathsf{E}[e^{j(\nu' X + \theta' Y)}]$ has the form

$$\exp\left\{ j(\nu' m_X + \theta' m_Y) \Leftrightarrow \tfrac{1}{2} \begin{bmatrix} \nu' & \theta' \end{bmatrix} \begin{bmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{bmatrix} \begin{bmatrix} \nu \\ \theta \end{bmatrix} \right\}. \tag{7.11}$$

Now observe that

$$\begin{bmatrix} \nu' & \theta' \end{bmatrix} \begin{bmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{bmatrix} \begin{bmatrix} \nu \\ \theta \end{bmatrix} \tag{7.12}$$

is equal to

$$\nu' R_X \nu + \theta' R_Y \theta + 2\theta' R_{YX} \nu.$$

On the other hand, if we put $w := \nu + j\theta$, and use (7.9), then (see Problem 26)

$$w^H C w = \nu'(R_X + R_Y)\nu + \theta'(R_X + R_Y)\theta + 2\theta'(R_{YX} \Leftrightarrow R_{XY})\nu.$$

Clearly, if

$$R_X \;=\; R_Y \quad \text{and} \quad R_{XY} \;=\; \Leftrightarrow R_{YX}, \tag{7.13}$$

then (7.12) is equal to $w^H C w/2$. Conversely, if (7.12) is equal to $w^H C w/2$ for all $w = \nu + j\theta$, then (7.13) holds (Problem 33). We say that a complex Gaussian random vector $Z = X + jY$ is **circularly symmetric** if (7.13) holds. If Z is circularly symmetric and zero mean, then its characteristic function is

$$\mathsf{E}[e^{j(\nu' X + \theta' Y)}] \;=\; e^{-w^H C w/4}, \quad w = \nu + j\theta.$$

The density corresponding to (7.11) is (assuming zero means)

$$f_{XY}(x,y) = \frac{\exp\left\{ \Leftrightarrow\frac{1}{2} \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right\}}{(2\pi)^n \sqrt{\det K}}, \tag{7.14}$$

where

$$K := \begin{bmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{bmatrix}.$$

It is shown in Problem 34 that under the assumption of circular symmetry (7.13),

$$f_{XY}(x,y) = \frac{e^{-z'C^{-1}z}}{\pi^n \det C}, \quad z = x + jy, \tag{7.15}$$

and that C is invertible if and only if K is invertible.

7.6. Notes

Notes §7.2: The Multivariate Gaussian Density

***Note* 1.** Recall that an $n \times n$ matrix R is symmetric if it is equal to its transpose; i.e., $R = R'$. It is positive definite if $c'Rc > 0$ for all $c \neq 0$. We show that the determinant of a positive-definite matrix is positive. A trivial modification of the derivation shows that the determinant of a positive semidefinite matrix is nonnegative. At the end of the note, we also define the square root of a positive-definite matrix.

We start with the well-known fact that a symmetric matrix can be diagonalized [22]; i.e., there is an $n \times n$ matrix P such that $P'P = PP' = I$ and such that $P'RP$ is a diagonal matrix, say

$$P'RP = \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}.$$

Next, from $P'RP = \Lambda$, we can easily obtain $R = P\Lambda P'$. Since the determinant of a product of matrices is the product of their determinants, $\det R = \det P \det \Lambda \det P'$. Since the determinants are numbers, they can be multiplied in any order. Thus,

$$\begin{aligned} \det R &= \det \Lambda \det P' \det P \\ &= \det \Lambda \det(P'P) \\ &= \det \Lambda \det I \\ &= \det \Lambda \\ &= \lambda_1 \cdots \lambda_n. \end{aligned}$$

Rewrite $P'RP = \Lambda$ as $RP = P\Lambda$. Then it is easy to see that the columns of P are eigenvectors of R; i.e., if P has columns $p_1, \ldots, p_n$, then $Rp_i = \lambda_i p_i$. Next, since $P'P = I$, each p_i satisfies $p_i' p_i = 1$. Since R is positive definite,

$$0 \;<\; p_i' R p_i \;=\; p_i'(\lambda_i p_i) \;=\; \lambda_i p_i' p_i \;=\; \lambda_i.$$

Thus, each eigenvalue $\lambda_i > 0$, and it follows that $\det R = \lambda_1 \cdots \lambda_n > 0$.

Because positive-definite matrices are diagonalizable with positive eigenvalues, it is easy to define their **square root** by

$$\sqrt{R} \;:=\; P \sqrt{\Lambda}\, P',$$

where

$$\sqrt{\Lambda} \;:=\; \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix}.$$

Thus, $\det \sqrt{R} = \sqrt{\lambda_1} \cdots \sqrt{\lambda_n} = \sqrt{\det R}$. Furthermore, from the definition of $\sqrt{R}$, it is clear that it is positive definite and satisfies $\sqrt{R}\sqrt{R} = R$. We also point out that since $R = P\Lambda P'$, $R^{-1} = P\Lambda^{-1}P'$, where Λ^{-1} is diagonal with diagonal entries $1/\lambda_i$; hence, $\sqrt{R^{-1}} = (\sqrt{R})^{-1}$. Finally, note that $\sqrt{R}\, R^{-1} \sqrt{R} = (P\sqrt{\Lambda}\, P')(P\Lambda^{-1}P')(P\sqrt{\Lambda}\, P') = I$.

7.7. Problems

Problems §7.1: Mean Vector, Covariance Matrix, and Characteristic Function

1. The input U to a certain amplifier is $N(0,1)$, and the output is $X = ZU + Y$, where the amplifier's random gain Z has density

   $$f_Z(z) \;=\; \tfrac{3}{7} z^2, \quad 1 \le z \le 2;$$

   and given $Z = z$, the amplifier's random bias Y is conditionally exponential with parameter z. Assuming that the input U is independent of the amplifier parameters Z and Y, find the mean vector and the covariance matrix of $[X, Y, Z]'$.

2. Find the mean vector and covariance matrix of $[X, Y, Z]'$ if

   $$f_{XYZ}(x,y,z) = \frac{2 \exp[\Leftrightarrow |x \Leftrightarrow y| \Leftrightarrow (y \Leftrightarrow z)^2/2]}{z^5 \sqrt{2\pi}}, \quad z \ge 1,$$

   and $f_{XYZ}(x,y,z) = 0$ otherwise.

3. Let X, Y, and Z be jointly continuous. Assume that $X \sim$ uniform$[1,2]$; that given $X = x$, $Y \sim \exp(1/x)$; and that given $X = x$ and $Y = y$, Z is $N(x,1)$. Find the mean vector and covariance matrix of $[X, Y, Z]'$.

4. Find the mean vector and covariance matrix of $[X, Y, Z]'$ if

$$f_{XYZ}(x, y, z) \;=\; \frac{e^{-(x-y)^2/2} e^{-(y-z)^2/2} e^{-z^2/2}}{(2\pi)^{3/2}}.$$

Also find the joint characteristic function of $[X, Y, Z]'$.

5. *Traces and Norms.*

(a) If M is a random $n \times n$ matrix, show that $\mathrm{tr}(\mathsf{E}[M]) = \mathsf{E}[\mathrm{tr}(M)]$.

(b) Let A and B be matrices of dimensions $m \times n$ and $n \times m$, respectively. Derive the formula $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. *Hint:* Recall that if $C = AB$, then

$$C_{ij} \;:=\; \sum_{k=1}^{n} A_{ik} B_{kj}.$$

(c) Show that if X is an n-dimensional random vector with covariance matrix R, then

$$\mathsf{E}[\|X \Leftrightarrow \mathsf{E}[X]\|^2] \;=\; \mathrm{tr}(R) \;=\; \sum_{i=1}^{n} \mathsf{var}(X_i).$$

(d) For $m \times n$ matrices U and V, show that

$$\mathrm{tr}(UV') \;=\; \sum_{i=1}^{m} \sum_{k=1}^{n} U_{ik} V_{ik}.$$

Show that if U is fixed and $\mathrm{tr}(UV') = 0$ for all matrices V, then $U = 0$.

Remark. What is going on here is that $\mathrm{tr}(UV')$ defines an **inner product** on the space of $m \times n$ matrices.

Problems §7.2: The Multivariate Gaussian

6. If X is a zero-mean, multivariate Gaussian random variable, show that

$$\mathsf{E}[(\nu' X X' \nu)^k] \;=\; (2k \Leftrightarrow 1)(2k \Leftrightarrow 3) \cdots 5 \cdot 3 \cdot 1 \cdot (\nu' \mathsf{E}[XX'] \nu)^k.$$

Hint: Example 3.6.

7. *Wick's Theorem.* Let $X \sim N(0, R)$ be n-dimensional. Let $(i_1, \ldots, i_{2k})$ be a vector of indices chosen from $\{1, \ldots, n\}$. Repetitions are allowed; e.g., $(1, 3, 3, 4)$. Derive **Wick's Theorem**,

$$\mathsf{E}[X_{i_1} \cdots X_{i_{2k}}] \;=\; \sum_{j_1, \ldots, j_{2k}} R_{j_1 j_2} \cdots R_{j_{2k-1} j_{2k}},$$

where the sum is over all $j_1, \ldots, j_{2k}$ that are permutations of $i_1, \ldots, i_{2k}$ *and* such that the product $R_{j_1 j_2} \cdots R_{j_{2k-1} j_{2k}}$ is distinct. *Hint:* The idea is to view both sides of the equation derived in the previous problem as a multivariate polynomial in the n variables $\nu_1, \ldots, \nu_n$. After collecting all terms on each side that involve $\nu_{i_1} \cdots \nu_{i_{2k}}$, the corresponding coefficients must be equal. In the expression

$$\begin{aligned} \mathsf{E}[(\nu' X)^{2k}] &= \mathsf{E}\left[\left(\sum_{j_1=1}^{n} \nu_{j_1} X_{j_1} \right) \cdots \left(\sum_{j_{2k}=1}^{n} \nu_{j_{2k}} X_{j_{2k}} \right) \right] \\ &= \sum_{j_1=1}^{n} \cdots \sum_{j_{2k}=1}^{n} \nu_{j_1} \cdots \nu_{j_{2k}} \mathsf{E}[X_{j_1} \cdots X_{j_{2k}}], \end{aligned}$$

we are only interested in those terms for which $j_1, \ldots, j_{2k}$ is a permutation of $i_1, \ldots, i_{2k}$. There are $(2k)!$ such terms, each equal to

$$\nu_{i_1} \cdots \nu_{i_{2k}} \mathsf{E}[X_{i_1} \cdots X_{i_{2k}}].$$

Similarly, from

$$(\nu' R \nu)^k = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i R_{ij} \nu_j \right)^k$$

we are only interested in terms of the form

$$\nu_{j_1} \nu_{j_2} \cdots \nu_{j_{2k-1}} \nu_{j_{2k}} R_{j_1 j_2} \cdots R_{j_{2k-1} j_{2k}},$$

where $j_1, \ldots, j_{2k}$ is a permutation of $i_1, \ldots, i_{2k}$. Now many of these permutations involve the same value of the product $R_{j_1 j_2} \cdots R_{j_{2k-1} j_{2k}}$. First, because R is symmetric, each factor R_{ij} also occurs as R_{ji}. This happens in 2^k different ways. Second, the order in which the R_{ij} are multiplied together occurs in $k!$ different ways.

8. Let X be a multivariate normal random vector with covariance matrix R. Use Wick's theorem of the previous problem to evaluate $\mathsf{E}[X_1 X_2 X_3 X_4]$, $\mathsf{E}[X_1 X_3^2 X_4]$, and $\mathsf{E}[X_1^2 X_2^2]$.

9. Evaluate (7.3) if $m = 0$ and

$$R = \begin{bmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 \end{bmatrix}.$$

Show that your result has the same form as the bivariate normal density in (5.13).

10. Let $X = [X_1, \ldots, X_n]' \sim N(m, R)$, and suppose that $Y = AX + b$, where A is a $p \times n$ matrix, and $b \in \mathbb{R}^p$. Find the mean and variance of Y.

11. Let $X_1, \ldots, X_n$ be i.i.d. $N(m, \sigma^2)$ random variables, and let $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ denote the average of the X_i. For $j = 1, \ldots, n$, put $Y_j := X_j \Leftrightarrow \overline{X}$. Show that $\mathsf{E}[Y_j] = 0$ and that $\mathsf{E}[\overline{X}Y_j] = 0$ for $j = 1, \ldots, n$.

12. Let $X = [X_1, \ldots, X_n]' \sim N(m, R)$, and suppose R is block diagonal, say

$$R \;=\; \begin{bmatrix} S & 0 \\ 0 & T \end{bmatrix},$$

where S and T are square submatrices with S being $s \times s$ and T being $t \times t$ with $s + t = n$. Put $U := [X_1, \ldots, X_s]'$ and $V := [X_{s+1}, \ldots, X_n]'$. Show that U and V are independent. *Hint:* Show that

$$\varphi_X(\nu) \;=\; \varphi_U(\nu_1, \ldots, \nu_s)\, \varphi_V(\nu_{s+1}, \ldots, \nu_n),$$

where φ_U is an s-variate normal characteristic function, and φ_V is a t-variate normal characteristic function. Use the notation $\alpha := [\nu_1, \ldots, \nu_s]'$ and $\beta := [\nu_{s+1}, \ldots, \nu_n]'$.

13. The digital signal processing chip in a wireless communication receiver generates the n-dimensional Gaussian vector X with mean zero and positive-definite covariance matrix R. It then computes the vector $Y = R^{-1/2}X$. (Since $R^{-1/2}$ is invertible, there is no loss of information in applying such a transformation.) Finally, the decision statistic $V = \|Y\|^2 := \sum_{k=1}^n Y_k^2$ is computed.

    (a) Find the multivariate density of Y.

    (b) Find the density of Y_k^2 for $k = 1, \ldots, n$.

    (c) Find the density of V.

Problems §7.3: Estimation of Random Vectors

14. Let X denote a random signal of known mean m_X and known covariance matrix R_X. Suppose that in order to estimate X, all we have available is the noisy measurement

$$Y \;=\; GX + W,$$

where G is a known gain matrix, and W is a noise vector with zero mean and known covariance matrix R_W. Further assume that the covariance between the signal and noise, R_{XW}, is zero. Find the linear MMSE estimate of X based on Y.

15. Let X and Y be random vectors with known means and covariance matrices. Do not assume zero means. Find the best *purely linear* estimate of X based on Y; i.e., find the matrix A that minimizes $\mathsf{E}[\|X \Leftrightarrow AY\|^2]$. Similarly, find the best constant estimate of X; i.e., find the vector b that minimizes $\mathsf{E}[\|X \Leftrightarrow b\|^2]$.

16. In this problem, you will show that $AR_Y = R_{XY}$ has a solution even if R_Y is singular. Recall that since R_Y is symmetric, it can be diagonalized [22]; i.e., there is an $n \times n$ matrix P such that $P'P = PP' = I$ and such that $P'R_YP = \Lambda$ is a diagonal matrix, say $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Define a new random variable $\tilde{Y} := P'Y$. Consider the problem of finding the best linear estimate of X based on $\tilde{Y}$. This leads to finding a matrix $\tilde{A}$ that solves

$$\tilde{A}R_{\tilde{Y}} \;=\; R_{X\tilde{Y}}. \tag{7.16}$$

    (a) Show that if $\tilde{A}$ solves (7.16), then $A := \tilde{A}P'$ solves $AR_Y = R_{XY}$.

    (b) Show that (7.16) has a solution. *Hint:* Use the fact that if $\mathsf{var}(\tilde{Y}_k) = 0$, then $\mathsf{cov}(X_i, \tilde{Y}_k) = 0$ as well. (This fact is an easy consequence of the Cauchy–Schwarz inequality, which is derived in Problem 1 of Chapter 6.)

17. Show that (7.8) implies (7.7).

18. Show that if $g_1(y)$ and $g_2(y)$ both solve (7.8), then $g_1 = g_2$ in the sense that

$$\mathsf{E}[|g_2(Y) \Leftrightarrow g_1(Y)|] \;=\; 0.$$

    Hint: Observe that

$$|g_2(y) \Leftrightarrow g_1(y)| \;=\; [g_2(y) \Leftrightarrow g_1(y)]I_{B_+}(y) \Leftrightarrow [g_2(y) \Leftrightarrow g_1(y)]I_{B_-}(y),$$

    where

$$B_+ \;:=\; \{y : g_2(y) > g_1(y)\} \quad \text{and} \quad B_- \;:=\; \{y : g_2(y) < g_1(y)\}.$$

    Now write down the four versions of (7.8) obtained with $g = g_2$ or $g = g_1$ and $h(y) = I_{B_+}(y)$ or $h(y) = I_{B_-}(y)$.

19. Write down the analogs of (7.7) and (7.8) when X is vector valued. Find $\mathsf{E}[X|Y = y]$ if $[X', Y']'$ is a Gaussian random vector such that X has mean m_X and covariance R_X, Y has mean m_Y and covariance R_Y, and $R_{XY} = \mathsf{cov}(X, Y)$.

20. Let X and Y be jointly normal random vectors as in the previous problem, and let the matrix A solve $AR_Y = R_{XY}$. Show that given $Y = y$, X is conditionally $N(m_X + A(y \Leftrightarrow m_Y), R_X \Leftrightarrow AR_{YX})$. *Hints:* First note that $(X \Leftrightarrow m_X) \Leftrightarrow A(Y \Leftrightarrow m_Y)$ and Y are uncorrelated and therefore independent by Problem 12. Next, observe that $\mathsf{E}[e^{j\nu'X}|Y = y]$ is equal to

$$\mathsf{E}\big[e^{j\nu'[(X-m_X)-A(Y-m_Y)]}e^{j\nu'[m_X+A(Y-m_Y)]}\big|Y = y\big].$$

Problems §7.4: Transformations of Random Vectors

21. Let X and Y have joint density $f_{XY}(x,y)$. Let $U := X+Y$ and $V := X \Leftrightarrow Y$. Find $f_{UV}(u,v)$.

22. Let X and Y be independent uniform$(0,1]$ random variables. Let $U := \sqrt{\Leftrightarrow 2\ln X}\cos(2\pi Y)$, and $V := \sqrt{\Leftrightarrow 2\ln Y}\sin(2\pi Y)$. Show that U and V are independent $N(0,1)$ random variables.

23. Let X and Y have joint density $f_{XY}(x,y)$. Let $U := X+Y$ and $V := X/(X+Y)$. Find $f_{UV}(u,v)$. Apply your result to the case where X and Y are independent gamma random variables $X \sim g_{p,\lambda}$ and $Y \sim g_{q,\lambda}$. What type of density is f_U? What type of density is f_V? *Hint:* Results from Problem 21(d) in Chapter 5 may be helpful.

Problems §7.5: Complex Random Variables and Vectors

24. Show that for a complex random variable $Z = X+jY$, $\mathsf{cov}(Z) = \mathsf{var}(X)+\mathsf{var}(Y)$.

25. Consider the complex random vector $Z = X+jY$ with covariance matrix C.

    (a) Show that $C = (R_X + R_Y) + j(R_{YX} \Leftrightarrow R_{XY})$.

    (b) If the circular symmetry conditions $R_X = R_Y$ and $R_{XY} = \Leftrightarrow R_{YX}$ hold, show that the diagonal elements of R_{XY} are zero; i.e., the components X_i and Y_i are uncorrelated.

    (c) If the circular symmetry conditions hold, and if C is a real matrix, show that X and Y are uncorrelated.

26. Let Z be a complex random vector with covariance matrix $C = R + jQ$ for real matrices R and Q.

    (a) Show that $R = R'$ and that $Q' = \Leftrightarrow Q$.

    (b) If $Q' = \Leftrightarrow Q$, show that $\nu' Q \nu = 0$.

    (c) If $w = \nu + j\theta$, show that

$$w^H C w \;=\; \nu' R \nu + \theta' R \theta + 2\theta' Q \nu.$$

27. Let $Z = X + jY$ be a complex random vector, and let $A = \alpha + j\beta$ be a complex matrix. Show that the transformation $Z \mapsto AZ$ is equivalent to

$$\begin{bmatrix} X \\ Y \end{bmatrix} \mapsto \begin{bmatrix} \alpha & \Leftrightarrow\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}.$$

    Hence, multiplying an n-dimensional complex random vector by an $n \times n$ complex matrix is a linear transformation of the $2n$-dimensional vector $[X', Y']'$. Now show that such a transformation preserves circular symmetry; i.e., if Z is circularly symmetric, then so is AZ.

28. Consider the complex random vector Θ partitioned as

$$\Theta \;=\; \begin{bmatrix} Z \\ W \end{bmatrix} \;=\; \begin{bmatrix} X + jY \\ U + jV \end{bmatrix},$$

where X, Y, U, and V are appropriately sized real random vectors. Since every complex random vector is identified with a real random vector of twice the length, it is convenient to put $\widetilde{Z} := [X', Y']'$ and $\widetilde{W} := [U', V']'$. Since the real and imaginary parts of Θ are $R := [X', U']'$ and $I := [Y', V']'$, we put

$$\widetilde{\Theta} \;:=\; \begin{bmatrix} R \\ I \end{bmatrix} \;=\; \begin{bmatrix} X \\ U \\ Y \\ V \end{bmatrix}.$$

Assume that Θ is Gaussian and circularly symmetric.

(a) Show that $C_{ZW} = 0$ if and only if $R_{\widetilde{Z}\widetilde{W}} = 0$.

(b) Show that the complex matrix $A = \alpha + j\beta$ solves $AC_W = C_{ZW}$ if and only if

$$\widetilde{A} \;:=\; \begin{bmatrix} \alpha & \Leftrightarrow\beta \\ \beta & \alpha \end{bmatrix}$$

solves $\widetilde{A}R_{\widetilde{W}} = R_{\widetilde{Z}\widetilde{W}}$.

(c) If A solves $AC_W = C_{ZW}$, show that given $W = w$, Z is conditionally Gaussian and circularly symmetric $N(m_Z + A(w \Leftrightarrow m_W), C_Z \Leftrightarrow AC_{WZ})$. *Hint:* Problem 20.

29. Let $Z = X + jY$ have density $f_Z(z) = e^{-|z|^2}/\pi$ as discussed in the text.

(a) Find $\mathsf{cov}(Z)$.

(b) Show that $2|Z|^2$ has a chi-squared density with 2 degrees of freedom.

30. Let $X \sim N(m_r, 1)$ and $Y \sim N(m_i, 1)$ be independent, and define the complex random variable $Z := X + jY$. Use the result of Problem 17 in Chapter 4 to show that $|Z|$ has the Rice density.

31. The base station of a wireless communication system generates an n-dimensional, complex, circularly symmetric, Gaussian random vector Z with mean zero and covariance matrix C. Let $W = C^{-1/2}Z$.

(a) Find the density of W.

(b) Let $W_k = U_k + jV_k$. Find the joint density of the pair of real random variables (U_k, V_k).

(c) If

$$\|W\|^2 \;:=\; \sum_{k=1}^{n} |W_k|^2 \;=\; \sum_{k=1}^{n} U_k^2 + V_k^2,$$

show that $2\|W\|^2$ has a chi-squared density with $2n$ degrees of freedom.

Remark. (*i*) The chi-squared density with $2n$ degrees of freedom is the same as the n-Erlang density, whose cdf has a closed-form expression given in Problem 12(c) in Chapter 3. (*ii*) By Problem 13 in Chapter 4, $\sqrt{2}\,\|W\|$ has a Nakagami-n density with parameter $\lambda = 1$.

32. Let M be a real symmetric matrix such that $u'Mu = 0$ for all real vectors u.

    (a) Show that $v'Mu = 0$ for all real vectors u and v. *Hint:* Consider the quantity $(u+v)'M(u+v)$.

    (b) Show that $M = 0$. *Hint:* Note that $M = 0$ if and only if $Mu = 0$ for all u, and $Mu = 0$ if and only if $\|Mu\| = 0$.

33. Show that if (7.12) is equal to $w^H C w$ for all $w = \nu + j\theta$, then (7.13) holds. *Hint:* Use the result of the preceding problem.

34. Assume that circular symmetry (7.13) holds. In this problem you will show that (7.14) reduces to (7.15).

    (a) Show that $\det K = (\det C)^2/2^{2n}$. *Hint:*

$$\begin{aligned}
\det(2K) &= \det\begin{bmatrix} 2R_X & \Leftrightarrow 2R_{YX} \\ 2R_{YX} & 2R_X \end{bmatrix} \\
&= \det\begin{bmatrix} 2R_X + j2R_{YX} & \Leftrightarrow 2R_{YX} \\ 2R_{YX} \Leftrightarrow j2R_X & 2R_X \end{bmatrix} \\
&= \det\begin{bmatrix} C & \Leftrightarrow 2R_{YX} \\ \Leftrightarrow jC & 2R_X \end{bmatrix} \\
&= \det\begin{bmatrix} C & \Leftrightarrow 2R_{YX} \\ 0 & C^H \end{bmatrix} = (\det C)^2.
\end{aligned}$$

    Remark. Thus, K is invertible if and only if C is invertible.

    (b) *Matrix Inverse Formula.* For any matrices A, B, C, and D, let $V = A + BCD$. If A and C are invertible, show that

$$V^{-1} = A^{-1} \Leftrightarrow A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

    by verifying that the formula for V^{-1} satisfies $VV^{-1} = I$.

    (c) Show that

$$K^{-1} = \begin{bmatrix} \Delta^{-1} & R_X^{-1}R_{YX}\Delta^{-1} \\ \Leftrightarrow \Delta^{-1}R_{YX}R_X^{-1} & \Delta^{-1} \end{bmatrix},$$

    where $\Delta := R_X + R_{YX}R_X^{-1}R_{YX}$, by verifying that $KK^{-1} = I$. *Hint:* Note that Δ^{-1} satisfies

$$\Delta^{-1} = R_X^{-1} \Leftrightarrow R_X^{-1}R_{YX}\Delta^{-1}R_{YX}R_X^{-1}.$$

(d) Show that $C^{-1} = (\Delta^{-1} \Leftrightarrow j R_X^{-1} R_{YX} \Delta^{-1})/2$ by verifying that $CC^{-1} = I$.

(e) Show that (7.14) reduces to (7.15). *Hint:* Before using the formula in part (d), first use the fact that $(\mathrm{Im}\, C^{-1})' = \Leftrightarrow \mathrm{Im}\, C^{-1}$; this fact can be seen using the equation for Δ^{-1} given in part (c).

CHAPTER 8

Advanced Concepts in Random Processes

The two most important continuous-time random processes are the Poisson process and the Wiener process, which are introduced in Sections 8.1 and 8.3, respectively. The construction of arbitrary random processes in discrete and continuous time using Kolmogorov's Theorem is discussed in Section 8.4.

In addition to the Poisson process, marked Poisson processes and shot noise are introduced in Section 8.1. The extension of the Poisson process to renewal processes is presented briefly in Section 8.2. In Section 8.3, the Wiener process is defined and then interpreted as integrated white noise. The Wiener integral is introduced. The approximation of the Wiener process via a random walk is also outlined. For random walks without finite second moments, it is shown by a simulation example that the limiting process is no longer a Wiener process.

8.1. The Poisson Process

A **counting process** $\{N_t, t \geq 0\}$ is a random process that counts how many times something happens from time zero up to and including time t. A sample path of such a process is shown in Figure 8.1. Such processes always have a

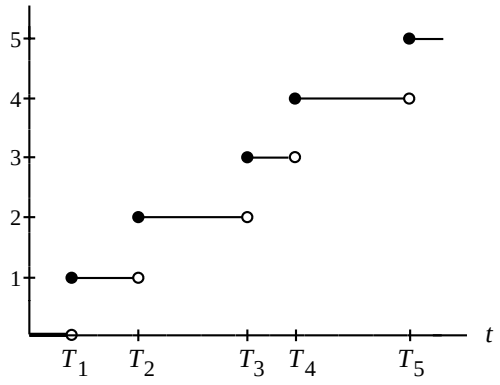

Figure 8.1. Sample path N_t of a counting process.

staircase form with jumps of height one. The randomness is in the times T_i at which whatever we are counting happens. Note that counting processes are right continuous.

Here are some examples of things that we might count:

- N_t = the number of radioactive particles emitted from a sample of radioactive material up to and including time t.

- N_t = the number of photoelectrons emitted from a photodetector up to and including time t.
- N_t = the number of hits of a web site up to and including time t.
- N_t = the number of customers passing through a checkout line at a grocery store up to and including time t.
- N_t = the number of vehicles passing through a toll booth on a highway up to and including time t.

Suppose that $0 \le t_1 < t_2 < \infty$ are given times, and we want to know how many things have happened between t_1 and t_2. Now N_{t_2} is the number of occurrences up to and including time t_2. If we subtract N_{t_1}, the number of occurrences up to and including time t_1, then the difference $N_{t_2} - N_{t_1}$ is simply the number of occurrences that happen after t_1 up to and including t_2. We call differences of the form $N_{t_2} - N_{t_1}$ **increments** of the process.

A counting process $\{N_t, t \ge 0\}$ is called a **Poisson process** if the following three conditions hold:

- $N_0 \equiv 0$; i.e., N_0 is a constant random variable whose value is always zero.
- For any $0 \le s < t < \infty$, the increment $N_t - N_s$ is a Poisson random variable with parameter $\lambda(t-s)$. The constant λ is called the **rate** or the **intensity** of the process.
- If the time intervals

  $$(t_1, t_2], (t_2, t_3], \ldots, (t_n, t_{n+1}]$$

  are disjoint, then the increments

  $$N_{t_2} - N_{t_1}, N_{t_3} - N_{t_2}, \ldots, N_{t_{n+1}} - N_{t_n}$$

  are independent; i.e., the process has **independent increments**. In other words, the numbers of occurrences in disjoint time intervals are independent.

***Example* 8.1.** Photoelectrons are emitted from a photodetector at a rate of λ per minute. Find the probability that during each of 2 consecutive minutes, more than 5 photoelectrons are emitted.

Solution. Let N_i denote the number of photoelectrons emitted from time zero up through the ith minute. The probability that during the first minute and during the second minute more than 5 photoelectrons are emitted is

$$\wp(\{N_1 - N_0 \ge 6\} \cap \{N_2 - N_1 \ge 6\}).$$

By the independent increments property, this is equal to

$$\wp(N_1 - N_0 \ge 6)\, \wp(N_2 - N_1 \ge 6).$$

Each of these factors is equal to

$$1 \Leftrightarrow \sum_{k=0}^{5} \frac{\lambda^k e^{-\lambda}}{k!},$$

where we have used the fact that the length of the time increments is one. Hence,

$$\wp(\{N_1 \Leftrightarrow N_0 \geq 6\} \cap \{N_2 \Leftrightarrow N_1 \geq 6\}) \;=\; \left(1 \Leftrightarrow \sum_{k=0}^{5} \frac{\lambda^k e^{-\lambda}}{k!}\right)^2.$$

We now compute the correlation and covariance functions of a Poisson process. Since $N_0 \equiv 0$, $N_t = N_t \Leftrightarrow N_0$ is a Poisson random variable with parameter $\lambda(t \Leftrightarrow 0) = \lambda t$. Hence,

$$\mathsf{E}[N_t] \;=\; \lambda t \quad \text{and} \quad \mathsf{var}(N_t) \;=\; \lambda t.$$

This further implies that $\mathsf{E}[N_t^2] = \lambda t + (\lambda t)^2$. For $0 \leq s < t$, we can compute the correlation

$$\begin{aligned} \mathsf{E}[N_t N_s] &= \mathsf{E}[(N_t \Leftrightarrow N_s) N_s] + \mathsf{E}[N_s^2] \\ &= \mathsf{E}[(N_t \Leftrightarrow N_s)(N_s \Leftrightarrow N_0)] + (\lambda s)^2 + \lambda s. \end{aligned}$$

Since $(0, s]$ and $(s, t]$ are disjoint, the above increments are independent, and so

$$\mathsf{E}[(N_t \Leftrightarrow N_s)(N_s \Leftrightarrow N_0)] = \mathsf{E}[N_t \Leftrightarrow N_s] \cdot \mathsf{E}[N_s \Leftrightarrow N_0] = \lambda(t \Leftrightarrow s) \cdot \lambda s.$$

It follows that

$$\mathsf{E}[N_t N_s] \;=\; (\lambda t)(\lambda s) + \lambda s.$$

We can also compute the covariance,

$$\begin{aligned} \mathsf{cov}(N_t, N_s) &= \mathsf{E}[(N_t \Leftrightarrow \lambda t)(N_s \Leftrightarrow \lambda s)] \\ &= \mathsf{E}[N_t N_s] \Leftrightarrow (\lambda t)(\lambda s) \\ &= \lambda s. \end{aligned}$$

More generally, given any two times t_1 and t_2,

$$\mathsf{cov}(N_{t_1}, N_{t_2}) \;=\; \lambda \min(t_1, t_2).$$

So far, we have focused on the *number* of occurrences between two fixed times. Now we focus on the **jump times**, which are defined by (see Figure 8.1)

$$T_n \;:=\; \min\{t > 0 : N_t \geq n\}.$$

In other words, T_n is the time of the nth jump in Figure 8.1. In particular, if $T_n > t$, then the nth jump happens after time t; hence, at time t we must

have $N_t < n$. Conversely, if at time t, $N_t < n$, then the nth occurrence has not happened yet; it must happen after time t, i.e., $T_n > t$. We can now write

$$\wp(T_n > t) \;=\; \wp(N_t < n) \;=\; \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

From Problem 12 in Chapter 3, we see that T_n has an Erlang density with parameters n and λ. In particular, T_1 has an exponential density with parameter λ. Depending on the context, the jump times are sometimes called **arrival times** or **occurrence times**.

In the previous paragraph, we defined the occurrence times in terms of counting process $\{N_t, t \geq 0\}$. Observe that we can express N_t in terms of the occurrence times since

$$N_t \;=\; \sum_{k=1}^{\infty} I_{(0,t]}(T_k).$$

We now define the **interarrival times**,

$$\begin{aligned} X_1 &= T_1, \\ X_n &= T_n \Leftrightarrow T_{n-1}, \quad n = 2, 3, \ldots. \end{aligned}$$

The occurrence times can be recovered from the interarrival times by writing

$$T_n \;=\; X_1 + \cdots + X_n.$$

We noted above that T_n is Erlang with parameters n and λ. Recalling Problem 46 in Chapter 3, which shows that a sum of i.i.d. exp(λ) random variables is Erlang with parameters n and λ, we wonder if the X_i are i.i.d. exponential with parameter λ. This is indeed the case, as shown in [4, p. 301].

***Example* 8.2.** Micrometeors strike the space shuttle according to a Poisson process. The expected time between strikes is 30 minutes. Find the probability that during at least one hour out of 5 consecutive hours, 3 or more micrometeors strike the shuttle.

Solution. The problem statement is telling us that the expected interarrival time is 30 minutes. Since the interarrival times are exp(λ) random variables, their mean is $1/\lambda$. Thus, $1/\lambda = 30$ minutes, or 0.5 hours, and so $\lambda = 2$ strikes per hour. The number of strikes during the ith hour is $N_i \Leftrightarrow N_{i-1}$. The probability that during at least one hour out of 5 consecutive hours, 3 or more micrometeors strike the shuttle is

$$\begin{aligned} \wp\Big(\bigcup_{i=1}^{5} \{N_i \Leftrightarrow N_{i-1} \geq 3\} \Big) &= 1 \Leftrightarrow \wp\Big(\bigcap_{i=1}^{5} \{N_i \Leftrightarrow N_{i-1} < 3\} \Big) \\ &= 1 \Leftrightarrow \prod_{i=1}^{5} \wp(N_i \Leftrightarrow N_{i-1} \leq 2), \end{aligned}$$

where the last step follows by the independent increments property of the Poisson process. Since $N_i \Leftrightarrow N_{i-1} \sim \text{Poisson}(\lambda[i \Leftrightarrow (i \Leftrightarrow 1)])$, or simply $\text{Poisson}(\lambda)$,

$$\wp(N_i \Leftrightarrow N_{i-1} \leq 2) \;=\; e^{-\lambda}(1+\lambda+\lambda^2/2) \;=\; 5e^{-2},$$

we have

$$\wp\left(\bigcup_{i=1}^{5}\{N_i \Leftrightarrow N_{i-1} \geq 3\}\right) \;=\; 1 \Leftrightarrow (5e^{-2})^5 \;\approx\; 0.86.$$

---

*Derivation of the Poisson Probabilities

In the definition of the Poisson process, we required that $N_t \Leftrightarrow N_s$ be a Poisson random variable with parameter $\lambda(t \Leftrightarrow s)$. In particular, this implies that

$$\frac{\wp(N_{t+\Delta t} \Leftrightarrow N_t = 1)}{\Delta t} \;=\; \frac{\lambda \Delta t\, e^{-\lambda \Delta t}}{\Delta t} \to \lambda,$$

as $\Delta t \to 0$. The Poisson assumption also implies that

$$\begin{aligned}
\frac{1 \Leftrightarrow \wp(N_{t+\Delta t} \Leftrightarrow N_t = 0)}{\Delta t} &= \frac{\wp(N_{t+\Delta t} \Leftrightarrow N_t \geq 1)}{\Delta t} \\
&= \frac{1}{\Delta t}\sum_{k=1}^{\infty}\frac{(\lambda\Delta t)^k}{k!} \\
&= \lambda\left(1+\sum_{k=2}^{\infty}\frac{(\lambda\Delta t)^{k-1}}{k!}\right) \to \lambda,
\end{aligned}$$

as $\Delta t \to 0$.

As we now show, the converse is also true as long as we continue to assume that $N_0 \equiv 0$ and that the process has independent increments. So, instead of assuming that $N_t \Leftrightarrow N_s$ is a Poisson random variable with parameter $\lambda(t \Leftrightarrow s)$, we assume that

- During a sufficiently short time interval, $\Delta t > 0$, $\wp(N_{t+\Delta t} \Leftrightarrow N_t = 1) \approx \lambda \Delta t$. By this we mean that

  $$\lim_{\Delta t \downarrow 0} \frac{\wp(N_{t+\Delta t} \Leftrightarrow N_t = 1)}{\Delta t} \;=\; \lambda.$$

  This property can be interpreted as saying that the probability of having exactly one occurrence during a short time interval of length Δt is approximately $\lambda \Delta t$.

- For sufficiently small $\Delta t > 0$, $\wp(N_{t+\Delta t} \Leftrightarrow N_t = 0) \approx 1 \Leftrightarrow \lambda \Delta t$. More precisely,

  $$\lim_{\Delta t \downarrow 0} \frac{1 \Leftrightarrow \wp(N_{t+\Delta t} \Leftrightarrow N_t = 0)}{\Delta t} \;=\; \lambda.$$

By combining this property with the preceding one, we see that during a short time interval of length Δt, we have either exactly one occurrence or no occurrences. In other words, during a short time interval, at most one occurrence is observed.

For $n = 0, 1, \ldots$, let

$$p_n(t) \;:=\; \wp(N_t \Leftrightarrow N_s = n), \qquad t \geq s. \tag{8.1}$$

Note that $p_0(s) = \wp(N_s \Leftrightarrow N_s = 0) = \wp(0 = 0) = \wp(\Omega) = 1$. Now,

$$\begin{aligned} p_n(t+\Delta t) &= \wp(N_{t+\Delta t} \Leftrightarrow N_s = n) \\ &= \wp\left(\bigcup_{k=0}^{n}\left[\{N_t \Leftrightarrow N_s = n \Leftrightarrow k\} \cap \{N_{t+\Delta t} \Leftrightarrow N_t = k\}\right]\right) \\ &= \sum_{k=0}^{n} \wp(N_{t+\Delta t} \Leftrightarrow N_t = k,\, N_t \Leftrightarrow N_s = n \Leftrightarrow k) \\ &= \sum_{k=0}^{n} \wp(N_{t+\Delta t} \Leftrightarrow N_t = k) p_{n-k}(t), \end{aligned}$$

using independent increments and (8.1). Break the preceding sum into three terms as follows.

$$\begin{aligned} p_n(t+\Delta t) &= \wp(N_{t+\Delta t} \Leftrightarrow N_t = 0) p_n(t) \\ &\quad + \wp(N_{t+\Delta t} \Leftrightarrow N_t = 1) p_{n-1}(t) \\ &\quad + \sum_{k=2}^{n} \wp(N_{t+\Delta t} \Leftrightarrow N_t = k) p_{n-k}(t). \end{aligned}$$

This enables us to write

$$\begin{aligned} p_n(t+\Delta t) \Leftrightarrow p_n(t) &= \Leftrightarrow[1 \Leftrightarrow \wp(N_{t+\Delta t} \Leftrightarrow N_t = 0)] p_n(t) \\ &\quad + \wp(N_{t+\Delta t} \Leftrightarrow N_t = 1) p_{n-1}(t) \\ &\quad + \sum_{k=2}^{n} \wp(N_{t+\Delta t} \Leftrightarrow N_t = k) p_{n-k}(t). \end{aligned} \tag{8.2}$$

For $n = 0$, only the first term on the right in (8.2) is present, and we can write

$$p_0(t+\Delta t) \Leftrightarrow p_0(t) \;=\; \Leftrightarrow[1 \Leftrightarrow \wp(N_{t+\Delta t} \Leftrightarrow N_t = 0)] p_0(t). \tag{8.3}$$

It then follows that

$$\lim_{\Delta t \downarrow 0} \frac{p_0(t+\Delta t) \Leftrightarrow p_0(t)}{\Delta t} \;=\; \Leftrightarrow \lambda p_0(t).$$

In other words, we are left with the first-order differential equation,

$$p_0'(t) \;=\; \Leftrightarrow \lambda p_0(t), \qquad p_0(s) = 1,$$

whose solution is simply

$$p_0(t) \;=\; e^{-\lambda(t-s)}, \qquad t \geq s.$$

To handle the case $n \geq 2$, note that since

$$\begin{aligned}
&\sum_{k=2}^{n} \wp(N_{t+\Delta t} - N_t = k) p_{n-k}(t) \\
&\qquad \leq \;\; \sum_{k=2}^{n} \wp(N_{t+\Delta t} - N_t = k) \\
&\qquad = \;\; \wp(N_{t+\Delta t} - N_t \geq 2) \\
&\qquad = \;\; 1 - [\, \wp(N_{t+\Delta t} - N_t = 0) + \wp(N_{t+\Delta t} - N_t = 1) \,],
\end{aligned}$$

it follows that

$$\lim_{\Delta t \downarrow 0} \frac{\sum_{k=2}^{n} \wp(N_{t+\Delta t} - N_t = k) p_{n-k}(t)}{\Delta t} \;=\; \lambda - \lambda \;=\; 0.$$

Returning to (8.2), we see that for $n = 1$ and for $n \geq 2$,

$$\lim_{\Delta t \downarrow 0} \frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} \;=\; -\lambda p_n(t) + \lambda p_{n-1}(t).$$

This results in the differential-difference equation,

$$p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \quad p_0(t) = e^{-\lambda(t-s)}. \tag{8.4}$$

It is easily verified that for $n = 1, 2, \ldots,$

$$p_n(t) \;=\; \frac{[\lambda(t-s)]^n e^{-\lambda(t-s)}}{n!},$$

which are the claimed Poisson probabilities, solve (8.4).

Marked Poisson Processes

It is frequently the case that in counting arrivals, each arrival is associated with a **mark**. For example, suppose packets arrive at a router according to a Poisson process of rate λ, and that the size of the ith packet is B_i bytes. The size B_i is the mark. Thus, the ith packet, whose size is B_i, arrives at time T_i, where T_i is the ith occurrence time of the Poisson process. The total number of bytes processed up to time t is

$$M_t \;:=\; \sum_{i=1}^{N_t} B_i.$$

We usually assume that the mark sequence is independent of the Poisson process. In this case, the mean of M_t can be computed as in Example 5.15. The characteristic function of M_t can be computed as in Problem 37 in Chapter 5.

Shot Noise

Light striking a photodetector generates photoelectrons according to a Poisson process. The rate of the process is proportional to the intensity of the light and the efficiency of the detector. The detector output is then passed through an amplifier of impulse response $h(t)$. We model the input to the amplifier as a train of impulses

$$X_t \;:=\; \sum_i \delta(t \Leftrightarrow T_i),$$

where the T_i are the occurrence times of the Poisson process. The amplifier output is

$$\begin{aligned} Y_t &= \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau) X_\tau \, d\tau \\ &= \sum_{i=1}^{\infty} \int_{-\infty}^{\infty} h(t \Leftrightarrow \tau) \delta(\tau \Leftrightarrow T_i) \, d\tau \\ &= \sum_{i=1}^{\infty} h(t \Leftrightarrow T_i). \end{aligned}$$

For any realizable system, $h(t)$ is a causal function; i.e., $h(t) = 0$ for $t < 0$. Then

$$Y_t \;=\; \sum_{i:T_i \le t} h(t \Leftrightarrow T_i) \;=\; \sum_{i=1}^{N_t} h(t \Leftrightarrow T_i). \tag{8.5}$$

A process of the form of Y_t is called a **shot-noise** process or a **filtered Poisson process**. Note that if the impulse response $h(t)$ is continuous, then so is Y_t. If $h(t)$ contains jump discontinuities, then so does Y_t. If $h(t)$ is nonnegative, then so is Y_t.

8.2. Renewal Processes

Recall that a Poisson process of rate λ can be constructed by writing

$$N_t \;:=\; \sum_{k=1}^{\infty} I_{[0,t]}(T_k),$$

where the arrival times

$$T_k \;:=\; X_1 + \cdots + X_k,$$

and the X_k are i.i.d. $\exp(\lambda)$ interarrival times. If we drop the requirement that the interarrival times be exponential and let them have arbitrary density f, then N_t is called a **renewal process**.

If we let F denote the cdf corresponding to the interarrival density f, it is easy to see that the mean of the process is

$$\mathsf{E}[N_t] \;=\; \sum_{k=1}^{\infty} F_k(t), \tag{8.6}$$

where F_k is the cdf of T_k. The corresponding density, denoted by f_k, is the k-fold convolution of f with itself. Hence, in general this formula is difficult to work with. However, there is another way to characterize $\mathsf{E}[N_t]$. In the problems you are asked to derive **renewal equation**

$$\mathsf{E}[N_t] \;=\; F(t) + \int_0^t \mathsf{E}[N_{t-x}] f(x)\, dx.$$

The mean function $m(t) := \mathsf{E}[N_t]$ of a renewal process is called the **renewal function**. Note that $m(0) = \mathsf{E}[N_0] = 0$, and that the renewal equation can be written in terms of the renewal function as

$$m(t) \;=\; F(t) + \int_0^t m(t \Leftrightarrow x) f(x)\, dx.$$

8.3. The Wiener Process

The **Wiener process** or **Brownian motion** is a random process that models integrated white noise. Such a model is needed because, as indicated below, white noise itself does not exist as an ordinary random process! In fact, the careful reader will note that in Chapter 6, we never worked directly with white noise, but rather with the output of an LTI system whose input was white noise. The Wiener process provides a mathematically well-defined object that can be used in modeling the output of an LTI system in response to (approximately) white noise.

We say that $\{W_t, t \geq 0\}$ is a Wiener process if the following four conditions hold:

- $W_0 \equiv 0$; i.e., W_0 is a constant random variable whose value is always zero.
- For any $0 \leq s \leq t < \infty$, the increment $W_t \Leftrightarrow W_s$ is a Gaussian random variable with zero mean and variance $\sigma^2 (t \Leftrightarrow s)$.
- If the time intervals

  $$(t_1, t_2], (t_2, t_3], \ldots, (t_n, t_{n+1}]$$

  are disjoint, then the increments

  $$W_{t_2} \Leftrightarrow W_{t_1}, W_{t_3} \Leftrightarrow W_{t_2}, \ldots, W_{t_{n+1}} \Leftrightarrow W_{t_n}$$

  are independent; i.e., the process has independent increments.
- For each sample point $\omega \in \Omega$, $W_t(\omega)$ as a function of t is continuous. More briefly, we just say that W_t has **continuous sample paths**.

Remarks. (i) If the parameter $\sigma^2 = 1$, then the process is called a **standard Wiener process**. A sample path of a standard Wiener process is shown in Figure 8.2.

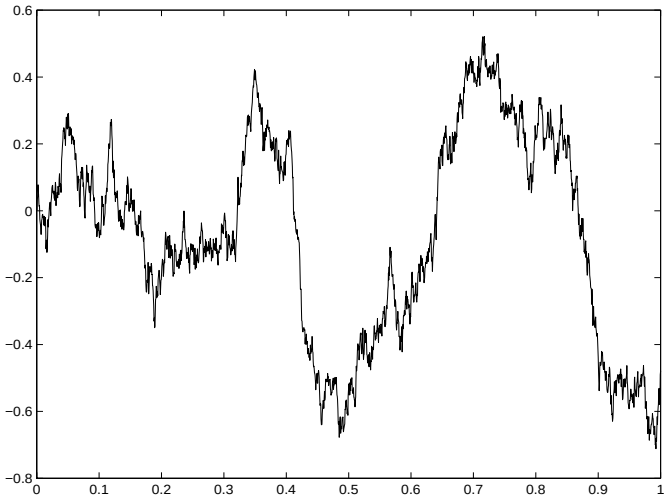

Figure 8.2. Sample path W_t of a standard Wiener process.

(*ii*) Since the first and third properties of the Wiener process are the same as those of the Poisson process, it is easy to show that

$$\mathsf{cov}(W_{t_1}, W_{t_2}) \;=\; \sigma^2 \min(t_1, t_2).$$

(*iii*) The fourth condition, that as a function of t, W_t should be continuous, is always assumed in practice, and can always be arranged by construction [4, Section 37]. Actually, the precise statement of the fourth property is

$$\wp(\{\omega \in \Omega : W_t(\omega) \text{ is a continuous function of } t\}) \;=\; 1,$$

i.e., the realizations of W_t are continuous with probability one. (*iv*) As indicated in Figure 8.2, the Wiener process is very "wiggly." In fact, it wiggles so much that it is nowhere differentiable with probability one [4, p. 505, Theorem 37.3].

Integrated White-Noise Interpretation of the Wiener Process

We now justify the interpretation of the Wiener process as integrated white noise. Let X_t be a zero-mean wide-sense stationary white noise process with correlation function $R_X(\tau) = \sigma^2\delta(\tau)$ and power spectral density $S_X(f) = \sigma^2$. For $t \geq 0$, put

$$W_t \;=\; \int_0^t X_\tau \, d\tau.$$

Then W_t is clearly zero mean. For $0 \leq s < t < \infty$, consider the increment

$$W_t \Leftrightarrow W_s \;=\; \int_0^t X_\tau \, d\tau \Leftrightarrow \int_0^s X_\tau \, d\tau \;=\; \int_s^t X_\tau \, d\tau.$$

This increment also has zero mean. To compute its variance, write

$$\begin{aligned}
\mathsf{var}(W_t \Leftrightarrow W_s) &= \mathsf{E}[(W_t \Leftrightarrow W_s)^2] \\
&= \mathsf{E}\bigg[\int_s^t X_\tau \, d\tau \int_s^t X_\theta \, d\theta\bigg] \\
&= \int_s^t \bigg(\int_s^t \mathsf{E}[X_\tau X_\theta] \, d\tau\bigg) d\theta \\
&= \int_s^t \bigg(\int_s^t \sigma^2 \delta(\tau \Leftrightarrow \theta) \, d\tau\bigg) d\theta \\
&= \int_s^t \sigma^2 \, d\theta \\
&= \sigma^2 (t \Leftrightarrow s). \qquad (8.7)
\end{aligned}$$

A random process X_t is said to be Gaussian (cf. Section 7.2), if for every function $c(\tau)$,

$$\int_{-\infty}^{\infty} c(\tau) X_\tau \, d\tau$$

is a Gaussian random variable. For example, if our white noise X_t is Gaussian, and if we take $c(\tau) = I_{(0,t]}(\tau)$, then

$$\int_{-\infty}^{\infty} c(\tau) X_\tau \, d\tau \;=\; \int_{-\infty}^{\infty} I_{(0,t]}(\tau) X_\tau \, d\tau \;=\; \int_0^t X_\tau \, d\tau \;=\; W_t$$

is a Gaussian random variable. Similarly,

$$W_t \Leftrightarrow W_s \;=\; \int_s^t X_\tau \, d\tau \;=\; \int_{-\infty}^{\infty} I_{(s,t]}(\tau) X_\tau \, d\tau$$

is a Gaussian random variable with zero mean and variance $\sigma^2(t \Leftrightarrow s)$. More generally, for time intervals

$$(t_1, t_2], (t_2, t_3], \ldots, (t_n, t_{n+1}],$$

consider the random vector

$$[W_{t_2} \Leftrightarrow W_{t_1}, W_{t_3} \Leftrightarrow W_{t_2}, \ldots, W_{t_{n+1}} \Leftrightarrow W_{t_n}]'.$$

Recall from Section 7.2 that this vector is normal, if every linear combination of its components is a Gaussian random variable. To see that this is indeed the case, write

$$\begin{aligned}
\sum_{i=1}^n c_i (W_{t_{i+1}} \Leftrightarrow W_{t_i}) &= \sum_{i=1}^n c_i \int_{t_i}^{t_{i+1}} X_\tau \, d\tau \\
&= \sum_{i=1}^n c_i \int_{-\infty}^{\infty} I_{(t_i, t_{i+1}]}(\tau) X_\tau \, d\tau \\
&= \int_{-\infty}^{\infty} \bigg(\sum_{i=1}^n c_i I_{(t_i, t_{i+1}]}(\tau)\bigg) X_\tau \, d\tau.
\end{aligned}$$

This is a Gaussian random variable if we assume X_t is a Gaussian random process. Finally, if the time intervals are disjoint, we claim that the increments are uncorrelated (and therefore independent under the Gaussian assumption). Without loss of generality, suppose $0 \leq s < t \leq \theta < \tau < \infty$. Write

$$\begin{aligned} \mathsf{E}[(W_t \Leftrightarrow W_s)(W_\tau \Leftrightarrow W_\theta)] &= \mathsf{E}\left[\int_s^t X_\alpha \, d\alpha \int_\theta^\tau X_\beta \, d\beta\right] \\ &= \int_\theta^\tau \left(\int_s^t \sigma^2 \delta(\alpha \Leftrightarrow \beta) \, d\alpha\right) d\beta. \end{aligned}$$

In the inside integral, we never have $\alpha = \beta$ because $s < \alpha \leq t \leq \theta < \beta \leq \tau$. Hence, the inner integral evaluates to zero, and the increments $W_t \Leftrightarrow W_s$ and $W_\tau \Leftrightarrow W_\theta$ are uncorrelated as claimed.

The Problem with White Noise

If white noise really existed and

$$W_t = \int_0^t X_\tau \, d\tau,$$

then the derivative of W_t would be $\dot{W}_t = X_t$. Now consider the following calculations. First, write

$$\begin{aligned} \frac{d}{dt}\mathsf{E}[W_t^2] &= \mathsf{E}\left[\frac{d}{dt} W_t^2\right] \\ &= \mathsf{E}[2W_t \dot{W}_t] \\ &= \mathsf{E}\left[2W_t \lim_{\Delta t \downarrow 0} \frac{W_{t+\Delta t} \Leftrightarrow W_t}{\Delta t}\right] \\ &= 2 \lim_{\Delta t \downarrow 0} \frac{\mathsf{E}[W_t(W_{t+\Delta t} \Leftrightarrow W_t)]}{\Delta t}. \end{aligned}$$

It is a simple calculation using the independent increments property to show that $\mathsf{E}[W_t(W_{t+\Delta t} \Leftrightarrow W_t)] = 0$. Hence

$$\frac{d}{dt}\mathsf{E}[W_t^2] = 0.$$

On the other hand, from (8.7), $\mathsf{E}[W_t^2] = \sigma^2 t$, and so

$$\frac{d}{dt}\mathsf{E}[W_t^2] = \frac{d}{dt}\sigma^2 t = \sigma^2 > 0.$$

The Wiener Integral

The Wiener process is a well-defined mathematical object. We argued above that W_t behaves like $\int_0^t X_\tau \, d\tau$, where X_t is white Gaussian noise. If such noise

is applied to an LTI system starting at time zero, and if the system has impulse response h, then the output is

$$\int_0^\infty h(t \Leftrightarrow \tau) X_\tau \, d\tau.$$

We now suppress t and write $g(\tau)$ instead of $h(t \Leftrightarrow \tau)$. Then we need a well-defined mathematical object to play the role of

$$\int_0^\infty g(\tau) X_\tau \, d\tau.$$

The required object is the **Wiener integral**,

$$\int_0^\infty g(\tau) \, dW_\tau \,,$$

which is defined as follows. For piecewise constant functions g of the form

$$g(\tau) \;=\; \sum_{i=1}^{n} g_i I_{(t_i, t_{i+1}]}(\tau),$$

where $0 \le t_1 < t_2 < \cdots < t_{n+1} < \infty$, we define

$$\int_0^\infty g(\tau) \, dW_\tau \;:=\; \sum_{i=1}^{n} g_i (W_{t_{i+1}} \Leftrightarrow W_{t_i}).$$

Note that the right-hand side is a weighted sum of independent, zero-mean, Gaussian random variables. The sum is therefore Gaussian with zero mean and variance

$$\sum_{i=1}^{n} g_i^2 \, \mathsf{var}(W_{t_{i+1}} \Leftrightarrow W_{t_i}) = \sum_{i=1}^{n} g_i^2 \cdot \sigma^2 (t_{i+1} \Leftrightarrow t_i) = \sigma^2 \int_0^\infty g(\tau)^2 \, d\tau.$$

Because of the zero mean, the variance and second moment are the same. Hence, we also have

$$\mathsf{E}\left[\left(\int_0^\infty g(\tau) \, dW_\tau \right)^2 \right] \;=\; \sigma^2 \int_0^\infty g(\tau)^2 \, d\tau. \tag{8.8}$$

For functions g that are not piecewise constant, but do satisfy $\int_0^\infty g(\tau)^2 \, d\tau < \infty$, the Wiener integral can be defined by a limiting process, which is discussed in more detail in Chapter 10. Basic properties of the Wiener integral are explored in the problems.

Random Walk Approximation of the Wiener Process

We now show how to approximate the Wiener process with a piecewise-constant, continuous-time process that is obtained from a discrete-time random walk.

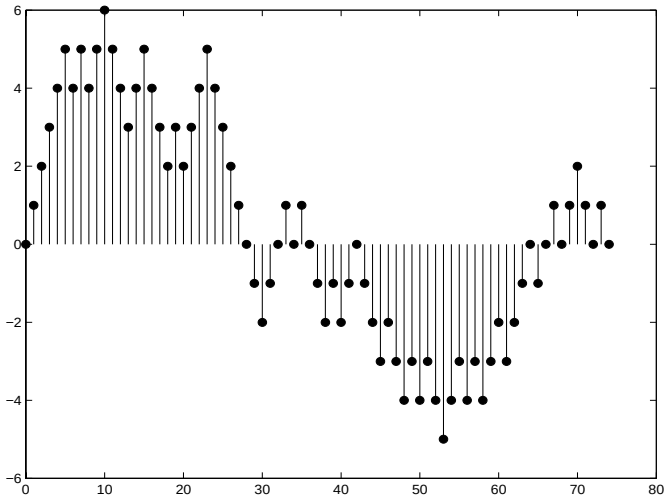

Figure 8.3. Sample path $S_k, k = 0, \ldots, 74$.

Let $X_1, X_2, \ldots$ be i.i.d. ± 1-valued random variables with $\wp(X_i = \pm 1) = 1/2$. Then each X_i has zero mean and variance one. Let

$$S_n \;:=\; \sum_{i=1}^{n} X_i .$$

Then S_n has zero mean and variance n. The process* $\{S_n, n \geq 0\}$ is called a symmetric random walk. A sample path of S_n is shown in Figure 8.3.

We next consider the scaled random walk $S_n/\sqrt{n}$. Note that $S_n/\sqrt{n}$ has zero mean and variance one. By the central limit theorem, which is discussed in detail in Chapter 4, the cdf of $S_n/\sqrt{n}$ converges to the standard normal cdf.

To approximate the continuous-time Wiener process, we use the piecewise-constant, continuous-time process

$$W_t^{(n)} \;:=\; \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor},$$

where $\lfloor \tau \rfloor$ denotes the greatest integer that is less than or equal to τ. For example, if $n = 100$ and $t = 3.1476$, then

$$W_{3.1476}^{(100)} \;=\; \frac{1}{\sqrt{100}} S_{\lfloor 100 \cdot 3.1476 \rfloor} \;=\; \frac{1}{10} S_{\lfloor 314.76 \rfloor} \;=\; \frac{1}{10} S_{314}.$$

Figure 8.4 shows a sample path of $W_t^{(75)}$ plotted as a function of t. As the continuous variable t ranges over $[0, 1]$, the values of $\lfloor 75t \rfloor$ range over the

*It is understood that $S_0 \equiv 0$.

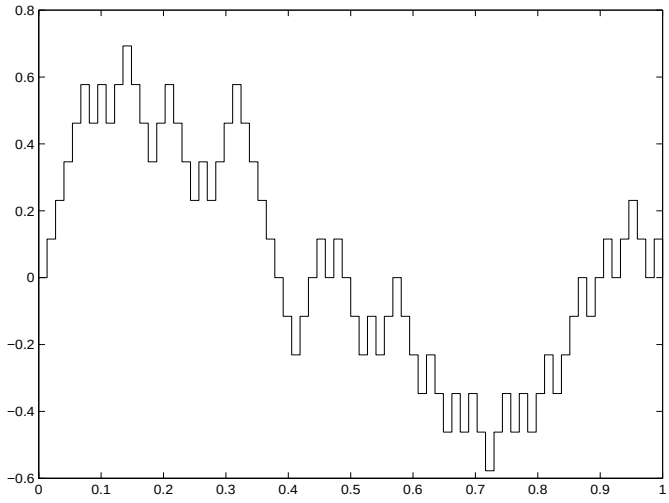

Figure 8.4. Sample path $W_t^{(75)}$.

integers $0, 1, \ldots, 75$. Thus, the constant levels seen in Figure 8.4 are

$$\frac{1}{\sqrt{75}} S_{\lfloor 75t \rfloor} \;=\; \frac{1}{\sqrt{75}} S_k, \quad k = 0, \ldots, 75.$$

In other words, the levels in Figure 8.4 are $1/\sqrt{75}$ times those in Figure 8.3.

Figures 8.5 and 8.6 show sample paths of $W_t^{(150)}$ and $W_t^{(10,000)}$, respectively. As n increases, the sample paths look more and more like those of the Wiener process shown in Figure 8.2.

Since the central limit theorem applies to any i.i.d. sequence with finite variance, the preceding convergence to the Wiener process holds if we replace the ± 1-valued X_i by any i.i.d. sequence with finite variance.† However, if the X_i only have finite mean but infinite variance, other limit processes can be obtained. For example, suppose the X_i are i.i.d. having a student's t density with $\nu = 3/2$ degrees of freedom. Then the X_i have zero mean and infinite variance. As can be seen in Figure 8.7, the limiting process has jumps, which is inconsistent with the Wiener process, which has continuous sample paths.

8.4. Specification of Random Processes

Finitely Many Random Variables

In this text we have often seen statements of the form, "Let X, Y, and Z be random variables with $\wp((X, Y, Z) \in B) = \mu(B)$," where $B \subset \mathbb{R}^3$, and $\mu(B)$

†If the mean of the X_i is m and the variance is σ^2, then we must replace S_n by $(S_n - nm)/\sigma$.

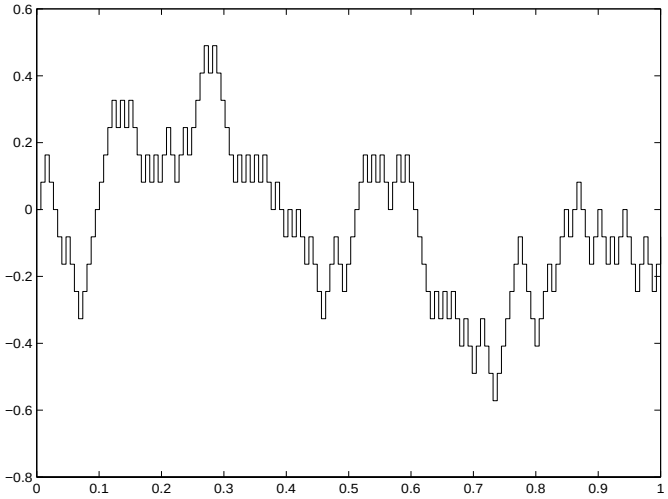

Figure 8.5. Sample path $W_t^{(150)}$.

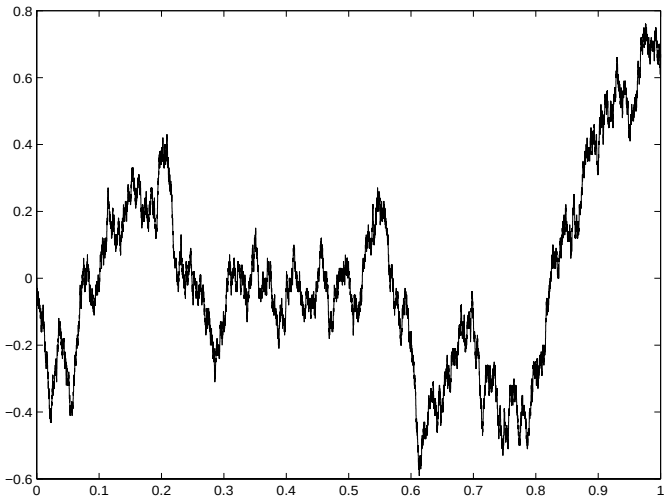

Figure 8.6. Sample path $W_t^{(10,000)}$.

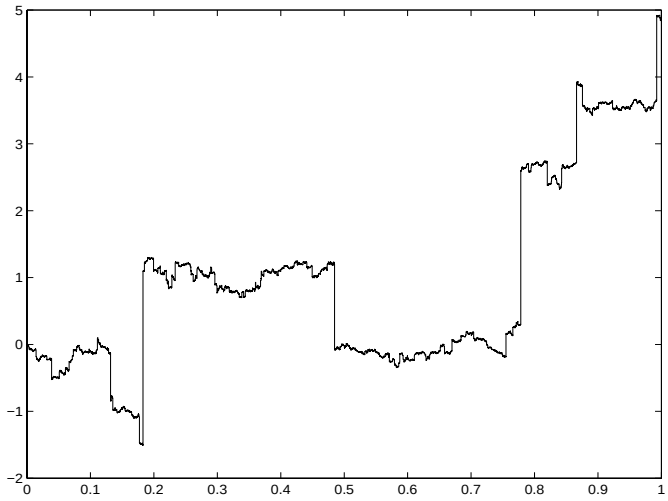

Figure 8.7. Sample path $S_{\lfloor nt \rfloor}/n^{2/3}$ for $n = 10,000$ when the X_i have a student's t density with $3/2$ degrees of freedom.

is given by some formula. For example, if X, Y, and Z are discrete, we would have

$$\mu(B) \;=\; \sum_i \sum_j \sum_k I_B(x_i, y_j, z_k) p_{i,j,k}, \tag{8.9}$$

where the x_i, y_j, and z_k are the values taken by the random variables, and the $p_{i,j,k}$ are nonnegative numbers that sum to one. If X, Y, and Z are jointly continuous, we would have

$$\mu(B) \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_B(x,y,z) f(x,y,z)\, dx\, dy\, dz, \tag{8.10}$$

where f is nonnegative and integrates to one. In fact, if X is discrete and Y and Z are jointly continuous, we would have

$$\mu(B) \;=\; \sum_i \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_B(x_i,y,z) f(x_i,y,z)\, dy\, dz, \tag{8.11}$$

where f is nonnegative and

$$\sum_i \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_i,y,z)\, dy\, dz \;=\; 1.$$

The big question is, given a formula for computing $\mu(B)$, how do we know that a sample space Ω, a probability measure $\wp$, and functions $X(\omega)$, $Y(\omega)$, and $Z(\omega)$ exist such that we indeed have

$$\wp\big((X,Y,Z) \in B\big) \;=\; \mu(B), \quad B \subset \mathbb{R}^3.$$

As we show in the next paragraph, the answer turns out to be rather simple.

If μ is defined by expressions such as (8.9)–(8.11), it can be shown that μ is a probability measure[‡] on $\mathbb{R}^3$; the case of (8.9) is easy; the other two require some background in measure theory, e.g., [4]. More generally, if we are given any probability measure μ on $\mathbb{R}^3$, we take $\Omega = \mathbb{R}^3$ and put $\wp(A) := \mu(A)$ for $A \subset \Omega = \mathbb{R}^3$. For $\omega = (\omega_1, \omega_2, \omega_3)$, we define

$$X(\omega) \;:=\; \omega_1, \quad Y(\omega) \;:=\; \omega_2, \quad \text{and} \quad Z(\omega) \;:=\; \omega_3.$$

It then follows that for $B \subset \mathbb{R}^3$,

$$\{\omega \in \Omega : (X(\omega), Y(\omega), Z(\omega)) \in B\}$$

reduces to

$$\{\omega \in \Omega : (\omega_1, \omega_2, \omega_3) \in B\} \;=\; B.$$

Hence, $\wp(\{(X, Y, Z) \in B\}) = \wp(B) = \mu(B)$.

For fixed $n \geq 1$, the foregoing ideas generalize in the obvious way to show the existence of a sample space Ω, probability measure $\wp$, and random variables $X_1, \ldots, X_n$ with

$$\wp\big((X_1, X_2, \ldots, X_n) \in B\big) \;=\; \mu(B), \quad B \subset \mathbb{R}^n,$$

where μ is any given probability measure defined on $\mathbb{R}^n$.

Infinite Sequences (Discrete Time)

Consider an infinite sequence of random variables such as $X_1, X_2, \ldots$. While $(X_1, \ldots, X_n)$ takes values in $\mathbb{R}^n$, the infinite sequence $(X_1, X_2, \ldots)$ takes values in $\mathbb{R}^\infty$. If such an infinite sequence of random variables exists on some sample space Ω equipped with some probability measure $\wp$, then[1]

$$\wp\big((X_1, X_2, \ldots) \in B\big), \quad B \subset \mathbb{R}^\infty,$$

is a probability measure on $\mathbb{R}^\infty$. We denote this probability measure by $\mu(B)$. Similarly, $\wp$ induces on $\mathbb{R}^n$ the measure

$$\mu_n(B_n) \;=\; \wp\big((X_1, \ldots, X_n) \in B_n\big), \quad B_n \subset \mathbb{R}^n,$$

Of course, $\wp$ induces on $\mathbb{R}^{n+1}$ the measure

$$\mu_{n+1}(B_{n+1}) \;=\; \wp\big((X_1, \ldots, X_n, X_{n+1}) \in B_{n+1}\big), \quad B_{n+1} \subset \mathbb{R}^{n+1}.$$

Now, if we take $B_{n+1} = B_n \times \mathbb{R}$ for any $B_n \subset \mathbb{R}^n$, then

$$\begin{aligned}
\mu_{n+1}(B_n \times \mathbb{R}) &= \wp\big((X_1, \ldots, X_n, X_{n+1}) \in B_n \times \mathbb{R}\big) \\
&= \wp\big((X_1, \ldots, X_n) \in B_n, X_{n+1} \in \mathbb{R}\big) \\
&= \wp\big(\{(X_1, \ldots, X_n) \in B_n\} \cap \{X_{n+1} \in \mathbb{R}\}\big) \\
&= \wp\big(\{(X_1, \ldots, X_n) \in B_n\} \cap \Omega\big) \\
&= \wp\big((X_1, \ldots, X_n) \in B_n\big) \\
&= \mu_n(B_n).
\end{aligned}$$

[‡]See Section 1.3 to review the definition of probability measure.

Thus, we have the **consistency condition**

$$\mu_{n+1}(B_n \times \mathbb{R}) \;=\; \mu_n(B_n), \quad B_n \subset \mathbb{R}^n, \quad n = 1, 2, \ldots. \tag{8.12}$$

Next, observe that since $(X_1, \ldots, X_n) \in B_n$ if and only if

$$(X_1, X_2, \ldots, X_n, X_{n+1}, \ldots) \in B_n \times \mathbb{R} \times \cdots,$$

it follows that $\mu_n(B_n)$ is equal to

$$\wp\big((X_1, X_2, \ldots, X_n, X_{n+1}, \ldots) \in B_n \times \mathbb{R} \times \cdots\big),$$

which is simply $\mu\big(B_n \times \mathbb{R} \times \cdots\big)$. Thus,

$$\mu(B_n \times \mathbb{R} \times \cdots) \;=\; \mu_n(B_n), \quad B_n \in \mathbb{R}^n, \quad n = 1, 2, \ldots. \tag{8.13}$$

The big question here is, if we are given a sequence of probability measures μ_n on $\mathbb{R}^n$ for $n = 1, 2, \ldots$, does there exist a probability measure μ on $\mathbb{R}^\infty$ such that (8.13) holds? The answer to this question was given by Kolmogorov, and is known as **Kolmogorov's Consistency Theorem** or as **Kolmogorov's Extension Theorem**. It says that if the consistency condition (8.12) holds,[§] then a probability measure μ exists on $\mathbb{R}^\infty$ such that (8.13) holds [8, p. 188].

We now specialize the foregoing discussion to the case of integer-valued random variables $X_1, X_2, \ldots$. For each $n = 1, 2, \ldots$, let $p_n(i_1, \ldots, i_n)$ denote a proposed joint probability mass function of $X_1, \ldots, X_n$. In other words, we want a random process for which

$$\wp\big((X_1, \ldots, X_n) \in B_n\big) \;=\; \sum_{i_1=-\infty}^{\infty} \cdots \sum_{i_n=-\infty}^{\infty} I_{B_n}(i_1, \ldots, i_n) p_n(i_1, \ldots, i_n).$$

More precisely, with $\mu_n(B_n)$ given by the above right-hand side, does there exist a measure μ on $\mathbb{R}^\infty$ such that (8.13) holds? By Kolmogorov's theorem, we just need to show that (8.12) holds.

We now show that (8.12) is equivalent to

$$\sum_{j=-\infty}^{\infty} p_{n+1}(i_1, \ldots, i_n, j) \;=\; p_n(i_1, \ldots, i_n). \tag{8.14}$$

[§]Knowing the measure μ_n, we can always write the corresponding cdf as

$$F_n(x_1, \ldots, x_n) \;=\; \mu_n\big((-\infty, x_1] \times \cdots \times (-\infty, x_n]\big).$$

Conversely, if we know the F_n, there is a unique measure μ_n on $\mathbb{R}^n$ such that the above formula holds [4, Section 12]. Hence, the consistency condition has the equivalent formulation in terms of cdfs [8, p. 189],

$$\lim_{x_{n+1} \to \infty} F_{n+1}(x_1, \ldots, x_n, x_{n+1}) \;=\; F_n(x_1, \ldots, x_n).$$

The left-hand side of (8.12) takes the form

$$\sum_{i_1=-\infty}^{\infty} \cdots \sum_{i_n=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} I_{B_n \times \mathbb{R}}(i_1, \ldots, i_n, j) p_{n+1}(i_1, \ldots, i_n, j). \tag{8.15}$$

Observe that $I_{B_n \times \mathbb{R}} = I_{B_n} I_{\mathbb{R}} = I_{B_n}$. Hence, the above sum becomes

$$\sum_{i_1=-\infty}^{\infty} \cdots \sum_{i_n=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} I_{B_n}(i_1, \ldots, i_n) p_{n+1}(i_1, \ldots, i_n, j),$$

which, using (8.14), simplifies to

$$\sum_{i_1=-\infty}^{\infty} \cdots \sum_{i_n=-\infty}^{\infty} I_{B_n}(i_1, \ldots, i_n) p_n(i_1, \ldots, i_n), \tag{8.16}$$

which is our definition of $\mu_n(B_n)$. Conversely, if in (8.12), or equivalently in (8.15) and (8.16), we take B_n to be the singleton set

$$B_n \;=\; \{(j_1, \ldots, j_n)\},$$

then we obtain (8.14).

The next question is how to construct a sequence of probability mass functions satisfying (8.14). Observe that (8.14) can be rewritten as

$$\sum_{j=-\infty}^{\infty} \frac{p_{n+1}(i_1, \ldots, i_n, j)}{p_n(i_1, \ldots, i_n)} \;=\; 1.$$

In other words, if $p_n(i_1, \ldots, i_n)$ is a valid joint pmf, and if we define

$$p_{n+1}(i_1, \ldots, i_n, j) \;:=\; p_{n+1|1,\ldots,n}(j|i_1, \ldots, i_n) \cdot p_n(i_1, \ldots, i_n),$$

where $p_{n+1|n,\ldots,1}(j|i_1, \ldots, i_n)$ is a valid pmf in the variable j (i.e., is nonnegative and the sum over j is one), then (8.14) will automatically hold!

***Example* 8.3.** Let $q(i)$ be any pmf. Take $p_1(i) := q(i)$, and take $p_{n+1|1,\ldots,n}(j|i_1, \ldots, i_n) := q(j)$. Then, for example,

$$p_2(i, j) \;=\; p_{2|1}(j|i)\, p_1(i) \;=\; q(j)\, q(i),$$

and

$$p_3(i, j, k) \;=\; p_{3|1,2}(k|i, j)\, p_2(i, j) \;=\; q(i)\, q(j)\, q(k).$$

More generally,

$$p_n(i_1, \ldots, i_n) \;=\; q(i_1) \cdots q(i_n).$$

Thus, the X_n are i.i.d. with common pmf q.

***Example* 8.4.** Again let $q(i)$ be any pmf. Suppose that for each i, $r(j|i)$ is a pmf in the variable j; i.e., r is any conditional pmf. Put $p_1(i) := q(i)$, and put[¶]

$$p_{n+1|1,\ldots,n}(j|i_1,\ldots,i_n) \;:=\; r(j|i_n).$$

Then

$$p_2(i,j) \;=\; p_{2|1}(j|i)\,p_1(i) \;=\; r(j|i)\,q(i),$$

and

$$p_3(i,j,k) \;=\; p_{3|1,2}(k|i,j)\,p_{1,2}(i,j) \;=\; q(i)\,r(j|i)\,r(k|j).$$

More generally,

$$p_n(i_1,\ldots,i_n) \;=\; q(i_1)\,r(i_2|i_1)\,r(i_3|i_2)\cdots r(i_n|i_{n-1}).$$

Continuous-Time Random Processes

The consistency condition for a continuous-time random process is a little more complicated. The reason is that in discrete time, between any two consecutive integers, there are no other integers, while in continuous time, for any $t_1 < t_2$, there are infinitely many times between t_1 and t_2.

Now suppose that for any $t_1 < \cdots < t_{n+1}$, we are given a probability measure $\mu_{t_1,\ldots,t_{n+1}}$ on $\mathbb{R}^{n+1}$. Fix any $B_n \subset \mathbb{R}^n$. For $k = 1,\ldots,n+1$, we define $B_{n,k} \subset \mathbb{R}^{n+1}$ by

$$B_{n,k} \;:=\; \{(x_1,\ldots,x_{n+1}) : (x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_{n+1}) \in B_n \text{ and } x_k \in \mathbb{R}\}.$$

Note the special cases[‖]

$$B_{n,1} \;=\; \mathbb{R}\times B_n \quad\text{and}\quad B_{n,n+1} \;=\; B_n \times \mathbb{R}.$$

The continuous-time consistency condition is that [40, p. 244] for $k = 1,\ldots,n+1$,

$$\mu_{t_1,\ldots,t_{n+1}}(B_{n,k}) = \mu_{t_1,\ldots,t_{k-1},t_{k+1},\ldots,t_{n+1}}(B_n). \tag{8.17}$$

If this condition holds, then there is a sample space Ω, a probability measure $\wp$, and random variables X_t such that

$$\wp\big((X_{t_1},\ldots,X_{t_n}) \in B_n\big) \;=\; \mu_{t_1,\ldots,t_n}(B_n), \quad B_n \subset \mathbb{R}^n,$$

for any $n \ge 1$ and any times $t_1 < \cdots < t_n$.

[¶] In Chapter 9, we will see that X_n is a Markov chain with stationary transition probabilities.

[‖] In the previous subsection, we only needed the case $k = n+1$. If we had wanted to allow two-sided discrete-time processes X_n for n any positive *or* negative integer, then both $k = 1$ and $k = n+1$ would have been needed (Problem 28).

8.5. Notes

Notes §8.4: Specification of Random Processes

***Note* 1.** Comments analogous to Note 1 in Chapter 5 apply here. Specifically, the set B must be restricted to a suitable σ-field $\mathcal{B}^\infty$ of subsets of $\mathbb{R}^\infty$. Typically, $\mathcal{B}^\infty$ is taken to be the smallest σ-field containing all sets of the form

$$\{\omega = (\omega_1, \omega_2, \ldots) \in \mathbb{R}^\infty : (\omega_1, \ldots, \omega_n) \in B_n\},$$

where B_n is a Borel subset of $\mathbb{R}^n$, and n ranges over the positive integers [4, p. 485].

8.6. Problems

Problems §8.1: The Poisson Process

1. Hits to a certain web site occur according to a Poisson process of rate $\lambda = 3$ per minute. What is the probability that there are no hits in a 10-minute period? Give a formula and then evaluate it to obtain a numerical answer.

2. Cell-phone calls processed by a certain wireless base station arrive according to a Poisson process of rate $\lambda = 12$ per minute. What is the probability that more than 3 calls arrive in a 20-second interval? Give a formula and then evaluate it to obtain a numerical answer.

3. Let N_t be a Poisson process with rate $\lambda = 2$, and consider a fixed observation interval $(0, 5]$.

   (a) What is the probability that $N_5 = 10$?

   (b) What is the probability that $N_i \Leftrightarrow N_{i-1} = 2$ for all $i = 1, \ldots, 5$?

4. A sports clothing store sells football jerseys with a certain very popular number on them according to a Poisson process of rate 3 crates per day. Find the probability that on 5 days in a row, the store sells at least 3 crates each day.

5. A sporting goods store sells a certain fishing rod according to a Poisson process of rate 2 per day. Find the probability that on at least one day during the week, the store sells at least 3 rods. (Note: week=five days.)

6. Let N_t be a Poisson process with rate λ, and let $\Delta t > 0$.

   (a) Show that $N_{t+\Delta t} \Leftrightarrow N_t$ and N_t are independent.

   (b) Show that $\wp(N_{t+\Delta t} = k + \ell | N_t = k) = \wp(N_{t+\Delta t} \Leftrightarrow N_t = \ell)$.

   (c) Evaluate $\wp(N_t = k | N_{t+\Delta t} = k + \ell)$.

   (d) Show that as a function of $k = 0, \ldots, n$, $\wp(N_t = k | N_{t+\Delta t} = n)$ has the binomial(n, p) probability mass function and identify p.

7. Customers arrive at a store according to a Poisson process of rate λ. What is the expected time until the nth customer arrives? What is the expected time between customers?

8. During the winter, snowstorms occur according to a Poisson process of intensity $\lambda = 2$ per week.

   (a) What is the average time between snowstorms?

   (b) What is the probability that no storms occur during a given two-week period?

   (c) If winter lasts 12 weeks, what is the expected number of snowstorms?

   (d) Find the probability that during at least one of the 12 weeks of winter, there are at least 5 snowstorms.

9. Space shuttles are launched according to a Poisson process. The average time between launches is two months.

   (a) Find the probability that there are no launches during a four-month period.

   (b) Find the probability that during at least one month out of four consecutive months, there are at least two launches.

10. Diners arrive at popular restaurant according to a Poisson process N_t of rate λ. A confused maitre d' seats the ith diner with probability p, and turns the diner away with probability $1 \Leftrightarrow p$. Let $Y_i = 1$ if the ith diner is seated, and $Y_i = 0$ otherwise. The number diners seated up to time t is

$$M_t \;:=\; \sum_{i=1}^{N_t} Y_i.$$

    Show that M_t is a Poisson random variable and find its parameter. Assume the Y_i are independent of each other and of the Poisson process.

    Remark. M_t is an example of a **thinned Poisson process**.

11. Lightning strikes occur according to a Poisson process of rate λ per minute. The energy of the ith strike is V_i. Assume the energy is independent of the occurrence times. What is the expected energy of a storm that lasts for t minutes? What is the average time between lightning strikes?

*12. Find the mean and characteristic function of the shot-noise random variable Y_t in equation (8.5).

Problems §8.2: Renewal Processes

13. In the case of a Poisson process, show that the right-hand side of (8.6) reduces to λt. *Hint:* Formula (4.2) in Section 4.2 may be helpful.

14. Derive the renewal equation

$$\mathsf{E}[N_t] \;=\; F(t) + \int_0^t \mathsf{E}[N_{t-x}] f(x)\, dx$$

as follows.

(a) Show that $\mathsf{E}[N_t | X_1 = x] = 0$ for $x > t$.

(b) Show that $\mathsf{E}[N_t | X_1 = x] = 1 + \mathsf{E}[N_{t-x}]$ for $x \leq t$.

(c) Use parts (a) and (b) and the laws of total probability and substitution to derive the renewal equation.

15. Solve the renewal equation for the renewal function $m(t) := \mathsf{E}[N_t]$ if the interarrival density is $f \sim \exp(\lambda)$. *Hint:* Differentiate the renewal equation and show that $m'(t) = \lambda$. It then follows that $m(t) = \lambda t$, which is what we expect since $f \sim \exp(\lambda)$ implies N_t is a Poisson process of rate λ.

Problems §8.3: The Wiener Process

16. For $0 \le s < t < \infty$, use the definition of the Wiener process to show that $\mathsf{E}[W_t W_s] = \sigma^2 s$.

17. Let the random vector $X = [W_{t_1}, \ldots, W_{t_n}]'$, $0 < t_1 < \cdots < t_n < \infty$, consist of samples of a Wiener process. Find the covariance matrix of X.

18. For piecewise constant g and h, show that

$$\int_0^\infty g(\tau)\, dW_\tau + \int_0^\infty h(\tau)\, dW_\tau \;=\; \int_0^\infty \big[g(\tau) + h(\tau)\big]\, dW_\tau .$$

Hint: The problem is easy if g and h are constant *over the same intervals*.

19. Use (8.8) to derive the formula

$$\mathsf{E}\left[\left(\int_0^\infty g(\tau)\, dW_\tau\right)\left(\int_0^\infty h(\tau)\, dW_\tau\right)\right] \;=\; \sigma^2 \int_0^\infty g(\tau)h(\tau)\, d\tau.$$

Hint: Consider the expectation

$$\mathsf{E}\left[\left(\int_0^\infty g(\tau)\, dW_\tau \Leftrightarrow \int_0^\infty h(\tau)\, dW_\tau\right)^2\right],$$

which can be evaluated in two different ways. The first way is to expand the square and take expectations term by term, applying (8.8) where possible. The second way is to observe that since

$$\int_0^\infty g(\tau)\, dW_\tau \Leftrightarrow \int_0^\infty h(\tau)\, dW_\tau \;=\; \int_0^\infty [g(\tau) \Leftrightarrow h(\tau)]\, dW_\tau ,$$

the above second moment can be computed directly using (8.8).

20. Let

$$Y_t \;=\; \int_0^t g(\tau)\, dW_\tau , \quad t \ge 0.$$

(a) Use (8.8) to show that

$$\mathsf{E}[Y_t^2] \;=\; \sigma^2 \int_0^t |g(\tau)|^2\, d\tau.$$

Hint: Observe that

$$\int_0^t g(\tau)\, dW_\tau \;=\; \int_0^\infty g(\tau) I_{(0,t]}(\tau)\, dW_\tau .$$

(b) Show that Y_t has correlation function

$$R_Y(t_1, t_2) \;=\; \sigma^2 \int_0^{\min(t_1,t_2)} |g(\tau)|^2\, d\tau, \quad t_1, t_2 \ge 0.$$

21. Consider the process

$$Y_t \;=\; e^{-\lambda t}V + \int_0^t e^{-\lambda(t-\tau)}\,dW_\tau, \quad t \geq 0,$$

where W_t is a Wiener process independent of V, and V has zero mean and variance q^2. Show that Y_t has correlation function

$$R_Y(t_1,t_2) \;=\; e^{-\lambda(t_1+t_2)}\Big(q^2 \Leftrightarrow \frac{\sigma^2}{2\lambda}\Big) + \frac{\sigma^2}{2\lambda}e^{-\lambda|t_1-t_2|}.$$

Remark. If V is normal, then the process Y_t is Gaussian and is known as an **Ornstein–Uhlenbeck process**.

22. Let W_t be a Wiener process, and put

$$Y_t \;:=\; \frac{e^{-\lambda t}}{\sqrt{2\lambda}}W_{e^{2\lambda t}}.$$

Show that

$$R_Y(t_1,t_2) \;=\; \frac{\sigma^2}{2\lambda}e^{-\lambda|t_1-t_2|}.$$

In light of the remark above, this is another way to define an Ornstein–Uhlenbeck process.

23. So far we have defined the Wiener process W_t only for $t \geq 0$. When defining W_t for all t, we continue to assume that $W_0 \equiv 0$; that for $s < t$, the increment $W_t \Leftrightarrow W_s$ is a Gaussian random variable with zero mean and variance $\sigma^2(t \Leftrightarrow s)$; that W_t has independent increments; and that W_t has continuous sample paths. The only difference is that s or both s and t can be negative, and that increments can be located anywhere in time, not just over intervals of positive time. In the following take $\sigma^2 = 1$.

(a) For $t > 0$, show that $\mathsf{E}[W_t^2] = t$.

(b) For $s < 0$, show that $\mathsf{E}[W_s^2] = \Leftrightarrow s$.

(c) Show that

$$\mathsf{E}[W_tW_s] \;=\; \frac{|t| + |s| \Leftrightarrow |t \Leftrightarrow s|}{2}.$$

Problems §8.4: Specification of Random Processes

24. Suppose X and Y are random variables with

$$\wp\big((X,Y) \in A\big) \;=\; \sum_i \int_{-\infty}^{\infty} I_A(x_i,y) f_{XY}(x_i,y)\,dy,$$

where the x_i are distinct real numbers, and f_{XY} is a nonnegative function satisfying

$$\sum_i \int_{-\infty}^{\infty} f_{XY}(x_i,y)\,dy \;=\; 1.$$

(a) Show that

$$\wp(X = x_k) \;=\; \int_{-\infty}^{\infty} f_{XY}(x_k, y)\, dy.$$

(b) Show that for $C \subset \mathbb{R}$,

$$\wp(Y \in C) \;=\; \int_C \left(\sum_i f_{XY}(x_i, y) \right) dy.$$

In other words, Y has marginal density

$$f_Y(y) \;=\; \sum_i f_{XY}(x_i, y).$$

(c) Show that

$$\wp(Y \in C | X = x_i) \;=\; \int_C \left[\frac{f_{XY}(x_i, y)}{p_X(x_i)} \right] dy.$$

In other words,

$$f_{Y|X}(y|x_i) \;=\; \frac{f_{XY}(x_i, y)}{p_X(x_i)}.$$

(d) For $B \subset \mathbb{R}$, show that if we define

$$\wp(X \in B | Y = y) \;:=\; \sum_i I_B(x_i) p_{X|Y}(x_i|y),$$

where

$$p_{X|Y}(x_i|y) \;:=\; \frac{f_{XY}(x_i, y)}{f_Y(y)},$$

then

$$\int_{-\infty}^{\infty} \wp(X \in B | Y = y) f_Y(y)\, dy \;=\; \wp(X \in B).$$

In other words, we have the law of total probability.

25. Let F be the standard normal cdf. Then F is a one-to-one mapping from $(\Leftrightarrow\infty, \infty)$ onto $(0, 1)$. Therefore, F has an inverse, $F^{-1}\colon (0, 1) \to (\Leftrightarrow\infty, \infty)$. If $U \sim \text{uniform}(0, 1)$, show that $X := F^{-1}(U)$ has F for its cdf.

26. Consider the cdf

$$F(x) \;:=\; \begin{cases} 0, & x < 0, \\ x^2, & 0 \le x < 1/2, \\ 1/4, & 1/2 \le x < 1, \\ x/2, & 1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

(a) Sketch $F(x)$.

(b) For $0 < u < 1$, sketch

$$G(u) \;:=\; \inf\{x \in \mathbb{R} : F(x) \ge u\}.$$

Hint: First identify the set

$$B_u \;:=\; \{x \in \mathbb{R} : F(x) \ge u\}.$$

Then find its infimum.

27. As illustrated in the previous problem, an arbitrary cdf F is usually not invertible, either because the equation $F(x) = u$ has more than one solution, e.g., $F(x) = 1/4$, or because it has no solution, e.g., $F(x) = 3/8$. However, for any cdf F, we can always introduce the function

$$G(u) \;:=\; \inf\{x \in \mathbb{R} : F(x) \ge u\}, \quad 0 < u < 1,$$

which, you will now show, can play the role of F^{-1} in Problem 25.

(a) Show that if $0 < u < 1$ and $x \in \mathbb{R}$, then $G(u) \le x$ if and only if $u \le F(x)$.

(b) Let $U \sim \text{uniform}(0,1)$, and put $X := G(U)$. Show that X has cdf F.

28. In the text we considered discrete-time processes X_n for $n = 1, 2, \ldots$. The consistency condition (8.12) arose from the requirement that

$$\wp\big((X_1, \ldots, X_n, X_{n+1}) \in B \times \mathbb{R}\big) \;=\; \wp\big((X_1, \ldots, X_n) \in B\big),$$

where $B \subset \mathbb{R}^n$. For processes X_n with $n = 0, \pm 1, \pm 2, \ldots$, we require not only

$$\wp\big((X_m, \ldots, X_n, X_{n+1}) \in B \times \mathbb{R}\big) \;=\; \wp\big((X_m, \ldots, X_n) \in B\big),$$

but also

$$\wp\big((X_{m-1}, X_m, \ldots, X_n) \in \mathbb{R} \times B\big) \;=\; \wp\big((X_m, \ldots, X_n) \in B\big),$$

where now $B \subset \mathbb{R}^{n-m+1}$. Let $\mu_{m,n}(B)$ be a proposed formula for the above right-hand side. Then the two consistency conditions are

$$\mu_{m,n+1}(B \times \mathbb{R}) \;= \mu_{m,n}(B) \quad \text{and} \quad \mu_{m-1,n}(\mathbb{R} \times B) \;=\; \mu_{m,n}(B).$$

For integer-valued random processes, show that these are equivalent to

$$\sum_{j=-\infty}^{\infty} p_{m,n+1}(i_m, \ldots, i_n, j) \;=\; p_{m,n}(i_m, \ldots, i_n)$$

and

$$\sum_{j=-\infty}^{\infty} p_{m-1,n}(j, i_m, \ldots, i_n) \;=\; p_{m,n}(i_m, \ldots, i_n),$$

where $p_{m,n}$ is the proposed joint probability mass function of $X_m, \ldots, X_n$.

29. Let q be any pmf, and let $r(j|i)$ be any conditional pmf. In addition, assume that $\sum_k q(k)r(j|k) = q(j)$. Put

$$p_{m,n}(i_m, \ldots, i_n) \;:=\; q(i_m)\, r(i_{m+1}|i_m)\, r(i_{m+2}|i_{m+1}) \cdots r(i_n|i_{n-1}).$$

Show that both consistency conditions for pmfs in the preceding problem are satisfied.

Remark. This process is strictly stationary as defined in Section 6.2 since the upon writing out the formula for $p_{m+k,n+k}(i_m, \ldots, i_n)$, we see that it does not depend on k.

30. Let μ_n be a probability measure on $\mathbb{R}^n$, and suppose that it is given in terms of a joint density f_n, i.e.,

$$\mu_n(B_n) \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_{B_n}(x_1, \ldots, x_n) f_n(x_1, \ldots, x_n)\, dx_n \cdots dx_1.$$

Show that the consistency condition (8.12) holds if and only if

$$\int_{-\infty}^{\infty} f_{n+1}(x_1, \ldots, x_n, x_{n+1})\, dx_{n+1} \;=\; f_n(x_1, \ldots, x_n).$$

31. Generalize the preceding problem for the continuous-time consistency condition (8.17).

32. Let W_t be a Wiener process. Let $f_{t_1,\ldots,t_n}$ denote the joint density of $W_{t_1}, \ldots, W_{t_n}$. Find $f_{t_1,\ldots,t_n}$ and show that it satisfies the density version of (8.17) that you derived in the preceding problem. *Hint:* The joint density should be Gaussian.

CHAPTER 9

Introduction to Markov Chains

A Markov chain is a random process with the property that given the values of the process from time zero up through the current time, the conditional probability of the value of the process at any future time depends only on its value at the current time. This is equivalent to saying that the future and the past are **conditionally independent** given the present (cf. Problem 29 in Chapter 1).

Markov chains often have intuitively pleasing interpretations. Some examples discussed in this chapter are random walks (without barriers and with barriers, which may be reflecting, absorbing, or neither), queuing systems (with finite or infinite buffers), birth–death processes (with or without spontaneous generation), life (with states being "healthy," "sick," and "death"), and the gambler's ruin problem.

Discrete-time Markov chains are introduced in Section 9.1 via the random walk and its variations. The emphasis is on the Chapman–Kolmogorov equation and the finding of stationary distributions.

Continuous-time Markov chains are introduced in Section 9.2 via the Poisson process. The emphasis is on Kolmogorov's forward and backward differential equations, and their use to find stationary distributions.

9.1. Discrete-Time Markov Chains

A sequence of discrete random variables, $X_0, X_1, \ldots$ is called a **Markov chain** if for $n \geq 1$,

$$\wp(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)$$

is equal to

$$\wp(X_{n+1} = x_{n+1} | X_n = x_n).$$

In other words, given the sequence of values $x_0, \ldots, x_n$, the conditional probability of what X_{n+1} will be depends only on the value of x_n.

***Example* 9.1** (Random Walk). Let X_0 be an integer-valued random variable that is independent of the i.i.d. sequence $Z_1, Z_2, \ldots$, where $\wp(Z_n = 1) = a$, $\wp(Z_n = \Leftrightarrow 1) = b$, and $\wp(Z_n = 0) = 1 \Leftrightarrow (a + b)$. Show that if

$$X_n \;:=\; X_{n-1} + Z_n, \quad n = 1, 2, \ldots,$$

then X_n is a Markov chain.

Solution. It helps to write out

$$X_1 \;=\; X_0 + Z_1$$

$$\begin{aligned} X_2 &= X_1 + Z_2 = X_0 + Z_1 + Z_2 \\ &\vdots \\ X_n &= X_{n-1} + Z_n = X_0 + Z_1 + \cdots + Z_n. \end{aligned}$$

The point here is that $(X_0, \ldots, X_n)$ is a function of $(X_0, Z_1, \ldots, Z_n)$, and hence, Z_{n+1} and $(X_0, \ldots, X_n)$ are independent. Now observe that

$$\wp(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0)$$

is equal to

$$\wp(X_n + Z_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0).$$

Using the substitution law, this becomes

$$\wp(Z_{n+1} = x_{n+1} \Leftrightarrow x_n | X_n = x_n, \ldots, X_0 = x_0).$$

On account of the independence of Z_{n+1} and $(X_0, \ldots, X_n)$, the above conditional probability is equal to

$$\wp(Z_{n+1} = x_{n+1} \Leftrightarrow x_n).$$

Since this depends on x_n but not on $x_{n-1}, \ldots, x_0$, it follows by the next example that $\wp(Z_{n+1} = x_{n+1} \Leftrightarrow x_n) = \wp(X_{n+1} = x_{n+1} | X_n = x_n)$.

The Markov chain of the preceding example is called a random walk on the integers. The random walk is said to be symmetric if $a = b = 1/2$. A realization of a symmetric random walk is shown in Figure 9.1. Notice that each point differs from the preceding one by ± 1.

To restrict the random walk to the nonnegative integers, we can take $X_n = \max(0, X_{n-1} + Z_n)$ (Problem 1).

***Example* 9.2.** Show that if

$$\wp(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0)$$

is a function of x_n but not $x_{n-1}, \ldots, x_0$, then the above conditional probability is equal to

$$\wp(X_{n+1} = x_{n+1} | X_n = x_n).$$

Solution. We use the identity $\wp(A \cap B) = \wp(A|B)\,\wp(B)$, where $A = \{X_{n+1} = x_{n+1}\}$ and $B = \{X_n = x_n, \ldots, X_0 = x_0\}$. Hence, if

$$\wp(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0) = h(x_n),$$

then the joint probability mass function

$$\wp(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)$$

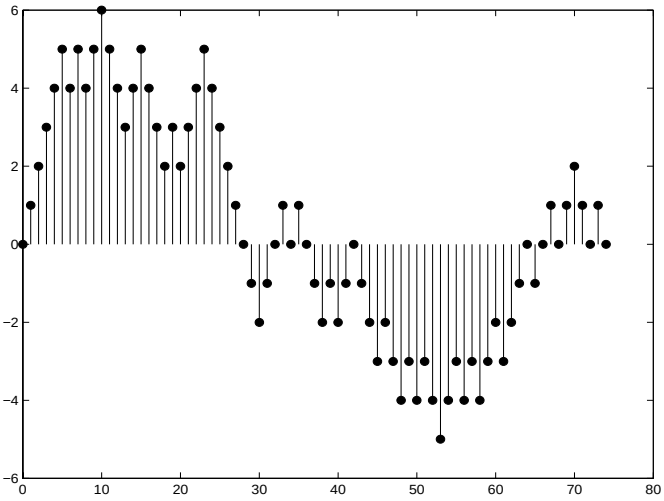

Figure 9.1. Realization of a symmetric random walk X_n.

is equal to

$$h(x_n)\, \wp(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0).$$

Summing both of these formulas over all values of $x_{n-1}, \ldots, x_0$ shows that

$$\wp(X_{n+1} = x_{n+1}, X_n = x_n) \;=\; h(x_n)\, \wp(X_n = x_n).$$

Hence, $h(x_n) = \wp(X_{n+1} = x_{n+1} | X_n = x_n)$ as claimed.

State Space and Transition Probabilities

The set of possible values that the random variables X_n can take is called the **state space** of the chain. In the rest of this chapter, we take the state space to be the set of integers or some specified subset of the integers. The conditional probabilities

$$\wp(X_{n+1} = j | X_n = i)$$

are called **transition probabilities**. In this chapter, we assume that the transition probabilities do not depend on time n. Such a Markov chain is said to have **stationary** transition probabilities or to be **time homogeneous**. For a time-homogeneous Markov chain, we use the notation

$$p_{ij} \;:=\; \wp(X_{n+1} = j | X_n = i)$$

for the transition probabilities. The p_{ij} are also called the **one-step transition probabilities** because they are the probabilities of going from state i to state j in one time step. One of the most common ways to specify the transition probabilities is with a **state transition diagram** as in Figure 9.2. This

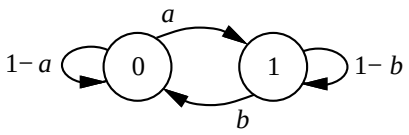

Figure 9.2. A state transition diagram. The diagram says that the state space is the finite set $\{0,1\}$, and that $p_{01} = a$, $p_{10} = b$, $p_{00} = 1 - a$, and $p_{11} = 1 - b$.

particular diagram says that the state space is the finite set $\{0,1\}$, and that $p_{01} = a$, $p_{10} = b$, $p_{00} = 1 \Leftrightarrow a$, and $p_{11} = 1 \Leftrightarrow b$. Note that the sum of all the probabilities leaving a state must be one. This is because for each state i,

$$\sum_j p_{ij} \;=\; \sum_j \wp(X_{n+1} = j | X_n = i) \;=\; 1.$$

The transition probabilities p_{ij} can be arranged in a matrix P, called the **transition matrix**, whose $i\,j$ entry is p_{ij}. For the chain in Figure 9.2,

$$P \;=\; \begin{bmatrix} 1 \Leftrightarrow a & a \\ b & 1 \Leftrightarrow b \end{bmatrix}.$$

The top row of P contains the probabilities p_{0j}, which is obtained by noting the probabilities written next to all the arrows leaving state 0. Similarly, the probabilities written next to all the arrows leaving state 1 are found in the bottom row of P.

Examples

The general **random walk** on the integers has the state transition diagram shown in Figure 9.3. Note that the Markov chain constructed in Example 9.1 is

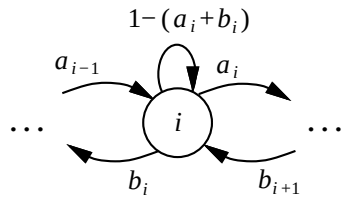

Figure 9.3. State transition diagram for a random walk on the integers.

a special case in which $a_i = a$ and $b_i = b$ for all i. The state transition diagram is telling us that

$$p_{ij} \;=\; \begin{cases} b_i, & j = i \Leftrightarrow 1, \\ 1 \Leftrightarrow (a_i + b_i), & j = i, \\ a_i, & j = i + 1, \\ 0, & \text{otherwise.} \end{cases} \tag{9.1}$$

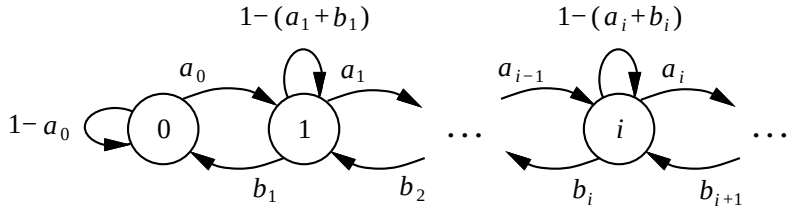

Figure 9.4. State transition diagram for a random walk with barrier at the origin (also called a birth–death process).

Hence, the transition matrix P is infinite, tridiagonal, and its ith row is

$$\left[\cdots \quad 0 \quad b_i \quad 1 \Leftrightarrow (a_i + b_i) \quad a_i \quad 0 \cdots \right].$$

Frequently, it is convenient to introduce a barrier at zero, leading to the state transition diagram in Figure 9.4. In this case, we speak of the random walk with barrier. If $a_0 = 1$, the barrier is said to be **reflecting**. If $a_0 = 0$, the barrier is said to be **absorbing**. Once a chain hits an absorbing state, the chain stays in that state from that time onward. If $a_i = a$ and $b_i = b$ for all i, the Markov chain is viewed as a model for a queue with an infinite buffer. A random walk with barrier at the origin is also known as a **birth–death process**. With this terminology, the state i is taken to be a population, say of bacteria. In this case, if $a_0 > 0$, there is **spontaneous generation**. If $b_i = 0$ for all i, we have a **pure birth process**. For $i \geq 1$, the formula for p_{ij} is given by (9.1) above, while for $i = 0$,

$$p_{0j} = \begin{cases} 1 \Leftrightarrow a_0, & j = 0, \\ a_0, & j = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{9.2}$$

The transition matrix P is the tridiagonal, semi-infinite matrix

$$P = \begin{bmatrix} 1 \Leftrightarrow a_0 & a_0 & 0 & 0 & 0 & \cdots \\ b_1 & 1 \Leftrightarrow (a_1 + b_1) & a_1 & 0 & 0 & \cdots \\ 0 & b_2 & 1 \Leftrightarrow (a_2 + b_2) & a_2 & 0 & \cdots \\ 0 & 0 & b_3 & 1 \Leftrightarrow (a_3 + b_3) & a_3 & \\ \vdots & \vdots & \vdots & & \ddots & \end{bmatrix}.$$

Sometimes it is useful to consider a random walk with barriers at the origin and at N, as shown in Figure 9.5. The formula for p_{ij} is given by (9.1) above for $1 \leq i \leq N \Leftrightarrow 1$, by (9.2) above for $i = 0$, and, for $i = N$, by

$$p_{Nj} = \begin{cases} b_N, & j = N \Leftrightarrow 1, \\ 1 \Leftrightarrow b_N, & j = N, \\ 0, & \text{otherwise.} \end{cases} \tag{9.3}$$

This chain is viewed as a model for a queue with a finite buffer, especially if $a_i = a$ and $b_i = b$ for all i. When $a_0 = 0$ and $b_N = 0$, the barriers at 0 and N

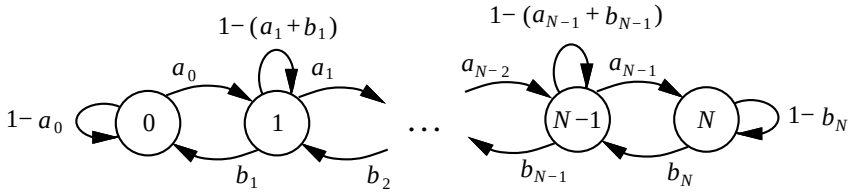

Figure 9.5. State transition diagram for a queue with a finite buffer.

are absorbing, and the chain is a model for the **gambler's ruin** problem. In this problem, a gambler starts at time zero with $1 \leq i \leq N \Leftrightarrow 1$ dollars, and plays until he either runs out of money, that is, absorption into state zero, or he acquires a fortune of N dollars and stops playing (absorption into state N). If $N = 2$ and $b_2 = 0$, the chain can be interpreted as the story of life if we view state $i = 0$ as being the "healthy" state, $i = 1$ as being the "sick" state, and $i = 2$ as being the "death" state. In this model, if you are healthy (in state 0), you remain healthy with probability $1 \Leftrightarrow a_0$ and become sick (move to state 1) with probability a_0. If you are sick (in state 1), you become healthy (move to state 0) with probability b_1, remain sick (stay in state 1) with probability $1 \Leftrightarrow (a_1 + b_1)$, or die (move to state 2) with probability b_1. Since state 2 is absorbing ($b_2 = 0$), once you enter this state, you never leave.

Stationary Distributions

The ***n*-step transition probabilities** are defined by

$$p_{ij}^{(n)} \;:=\; \wp(X_n = j | X_0 = i).$$

This is the probability of going from state i (at time zero) to state j in n steps. For a chain with stationary transition probabilities, it will be shown later that the n-step transition probabilities are also stationary; i.e., for $m = 1, 2, \ldots$,

$$\wp(X_{n+m} = j | X_m = i) \;=\; \wp(X_n = j | X_0 = i) \;=\; p_{ij}^{(n)}.$$

We also note that

$$p_{ij}^{(0)} \;=\; \wp(X_0 = j | X_0 = i) \;=\; \delta_{ij},$$

where δ_{ij} denotes the **Kronecker delta**, which is one if $i = j$ and is zero otherwise.

A Markov chain with stationary transition probabilities also satisfies the **Chapman–Kolmogorov equation**,

$$p_{ij}^{(n+m)} \;=\; \sum_k p_{ik}^{(n)}\, p_{kj}^{(m)}. \tag{9.4}$$

This equation, which is derived later, says that to go from state i to state j in $n + m$ steps, first you go to some intermediate state k in n steps, and then you

go from state k to state j in m steps. Taking $m = 1$, we have the special case

$$p_{ij}^{(n+1)} \;=\; \sum_k p_{ik}^{(n)}\, p_{kj}.$$

If $n = 1$ as well,

$$p_{ij}^{(2)} \;=\; \sum_k p_{ik}\, p_{kj}.$$

This equation says that the matrix with entries $p_{ij}^{(2)}$ is equal to the product PP. More generally, the matrix with entries $p_{ij}^{(n)}$, called the ***n*-step transition matrix** , is equal to P^n. The Chapman–Kolmogorov equation thus says that

$$P^{n+m} \;=\; P^n P^m.$$

For many Markov chains, it can be shown [17, Section 6.4] that

$$\lim_{n\to\infty} \wp(X_n = j | X_0 = i)$$

exists and does not depend on i. In other words, if the chain runs for a long time, it reaches a "steady state" in which the probability of being in state j does not depend on the initial state of the chain. If the above limit exists and does not depend on i, we put

$$\pi_j \;:=\; \lim_{n\to\infty} \wp(X_n = j | X_0 = i).$$

We call $\{\pi_j\}$ the **stationary distribution** or the **equilibrium distribution** of the chain. Note that if we sum both sides over j, we get

$$\begin{aligned} \sum_j \pi_j \;&=\; \sum_j \lim_{n\to\infty} \wp(X_n = j | X_0 = i) \\ &=\; \lim_{n\to\infty} \sum_j \wp(X_n = j | X_0 = i) \\ &=\; \lim_{n\to\infty} 1 \;=\; 1. \end{aligned} \tag{9.5}$$

To find the π_j, we can use the Chapman–Kolmogorov equation. If $p_{ij}^{(n)} \to \pi_j$, then taking limits in

$$p_{ij}^{(n+1)} \;=\; \sum_k p_{ik}^{(n)}\, p_{kj}$$

shows that

$$\pi_j \;=\; \sum_k \pi_k\, p_{kj}.$$

If we think of π as a row vector with entries π_j, and if we think of P as the matrix with entries p_{ij}, then the above equation says that

$$\pi \;=\; \pi P.$$

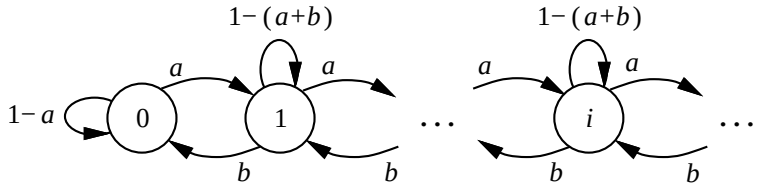

Figure 9.6. State transition diagram for a queuing system with an infinite buffer.

On account of (9.5), π cannot be the zero vector. Hence, π is a left **eigenvector** of P with **eigenvalue** 1. To say this another way, $I \Leftrightarrow P$ is singular; i.e., there are many solutions of $\pi(I \Leftrightarrow P) = 0$. The solution we want must satisfy not only $\pi = \pi P$, but also the normalization condition,

$$\sum_j \pi_j \;=\; 1,$$

derived in (9.5).

***Example* 9.3.** The state transition diagram for a queuing system with an infinite buffer is shown in Figure 9.6. Find the stationary distribution of the chain.

Solution. We begin by writing out

$$\pi_j \;=\; \sum_k \pi_k \, p_{kj} \tag{9.6}$$

for $j = 0, 1, 2, \ldots$. For each j, the coefficients p_{kj} are obtained by inspection of the state transition diagram. For $j = 0$, we must consider

$$\pi_0 \;=\; \sum_k \pi_k \, p_{k0}.$$

We need the values of p_{k0}. From the diagram, the only way to get to state 0 is from state 0 itself (with probability $p_{00} = 1 \Leftrightarrow a$) or from state 1 (with probability $p_{10} = b$). The other $p_{k0} = 0$. Hence,

$$\pi_0 \;=\; \pi_0 \, (1 \Leftrightarrow a) + \pi_1 \, b.$$

We can rearrange this to get

$$\pi_1 \;=\; \frac{a}{b} \pi_0.$$

Now put $j = 1$ in (9.6). The state transition diagram tells us that the only way to enter state 1 is from states 0, 1, and 2, with probabilities a, $1 \Leftrightarrow (a + b)$, and b, respectively. Hence,

$$\pi_1 \;=\; \pi_0 \, a + \pi_1 \, [1 \Leftrightarrow (a + b)] + \pi_2 \, b.$$

Substituting $\pi_1 = (a/b)\pi_0$ yields $\pi_2 = (a/b)^2\pi_0$. In general, if we substitute $\pi_j = (a/b)^j\pi_0$ and $\pi_{j-1} = (a/b)^{j-1}\pi_0$ into

$$\pi_j \;=\; \pi_{j-1}\,a + \pi_j\,[1 \Leftrightarrow (a+b)] + \pi_{j+1}\,b,$$

then we obtain $\pi_{j+1} = (a/b)^{j+1}\pi_0$. We conclude that

$$\pi_j \;=\; \left(\frac{a}{b}\right)^j \pi_0, \quad j = 0, 1, 2, \ldots.$$

To solve for π_0, we use the fact that

$$\sum_{j=0}^{\infty} \pi_j \;=\; 1,$$

or

$$\pi_0 \sum_{j=0}^{\infty} \left(\frac{a}{b}\right)^j \;=\; 1.$$

The geometric series formula shows that

$$\pi_0 \;=\; 1 \Leftrightarrow a/b,$$

and

$$\pi_j \;=\; \left(\frac{a}{b}\right)^j (1 \Leftrightarrow a/b).$$

In other words, the stationary distribution is a geometric$_0(a/b)$ probability mass function.

Derivation of the Chapman–Kolmogorov Equation

We begin by showing that

$$p_{ij}^{(n+1)} \;=\; \sum_l p_{il}^{(n)}\, p_{lj}\,.$$

To derive this, we need the following **law of total conditional probability**. We claim that

$$\wp(X = x|Z = z) \;=\; \sum_y \wp(X = x|Y = y, Z = z)\,\wp(Y = y|Z = z).$$

The right-hand side of this equation is simply

$$\sum_y \frac{\wp(X = x, Y = y, Z = z)}{\wp(Y = y, Z = z)} \cdot \frac{\wp(Y = y, Z = z)}{\wp(Z = z)}.$$

Canceling common factors and then summing over y yields

$$\frac{\wp(X = x, Z = z)}{\wp(Z = z)} \;=\; \wp(X = x|Z = z).$$

We can now write

$$\begin{aligned} p_{ij}^{(n+1)} &= \wp(X_{n+1}=j|X_0=i) \\ &= \sum_{l_n}\cdots\sum_{l_1}\wp(X_{n+1}=j|X_n=l_n,\ldots,X_1=l_1,X_0=i) \\ &\qquad\cdot\wp(X_n=l_n,\ldots,X_1=l_1|X_0=i) \\ &= \sum_{l_n}\cdots\sum_{l_1}\wp(X_{n+1}=j|X_n=l_n)\,\wp(X_n=l_n,\ldots,X_1=l_1|X_0=i) \\ &= \sum_{l_n}\wp(X_{n+1}=j|X_n=l_n)\,\wp(X_n=l_n|X_0=i) \\ &= \sum_{l} p_{il}^{(n)}p_{lj}\,. \end{aligned}$$

This establishes that for $m=1$,

$$p_{ij}^{(n+m)} = \sum_k p_{ik}^{(n)}p_{kj}^{(m)}\,.$$

We now assume it is true for m and show it is true for $m+1$. Write

$$\begin{aligned} p_{ij}^{(n+[m+1])} &= p_{ij}^{([n+m]+1)} \\ &= \sum_l p_{il}^{(n+m)}p_{lj} \\ &= \sum_l\left(\sum_k p_{ik}^{(n)}p_{kl}^{(m)}\right)p_{lj} \\ &= \sum_k p_{ik}^{(n)}\left(\sum_l p_{kl}^{(m)}p_{lj}\right) \\ &= \sum_k p_{ik}^{(n)}p_{kj}^{(m+1)}\,. \end{aligned}$$

Stationarity of the n-step Transition Probabilities

We would now like to use the Chapman–Kolmogorov equation to show that the n-step transition probabilities are stationary; i.e.,

$$\wp(X_{n+m}=j|X_m=i) = \wp(X_n=j|X_0=i) = p_{ij}^{(n)}.$$

To do this, we need the fact that for any $0\le\nu<n$,

$$\wp(X_{n+1}=j|X_n=i,X_\nu=l) = \wp(X_{n+1}=j|X_n=i). \tag{9.7}$$

To establish this fact, we use the law of total conditional probability to write the left-hand side as

$$\sum_{l_{n-1}}\cdots\sum_{l_0}\wp(X_{n+1}=j|X_n=i,X_{n-1}=l_{n-1},\ldots,X_\nu=l,\ldots,X_0=l_0)$$

$$\cdot\wp(X_{n-1}=l_{n-1},\ldots,X_{\nu+1}=l_{\nu+1},X_{\nu-1}=l_{\nu-1},\ldots,X_0=l_0|X_n=i,X_\nu=l),$$

where the sum is $(n \Leftrightarrow 1)$-fold, the sum over l_ν being omitted. By the Markov property, this simplifies to

$$\sum_{l_{n-1}} \cdots \sum_{l_0} \wp(X_{n+1} = j | X_n = i)$$

$$\cdot \wp(X_{n-1} = l_{n-1}, \ldots, X_{\nu+1} = l_{\nu+1}, X_{\nu-1} = l_{\nu-1}, \ldots, X_0 = l_0 | X_n = i, X_\nu = l),$$

which further simplifies to $\wp(X_{n+1} = j | X_n = i)$.

We can now show that the n-step transition probabilities are stationary. For a time-homogeneous Markov chain, the case $n = 1$ holds by definition. Assume it holds for m, and use the law of total conditional probability to write

$$\begin{aligned}
\wp(X_{n+m+1} = j | X_m = i) &= \sum_k \wp(X_{n+m+1} = j | X_{n+m} = k, X_m = i) \\
&\qquad \cdot \wp(X_{n+m} = k | X_m = i) \\
&= \sum_k \wp(X_{n+m+1} = j | X_{n+m} = k) \\
&\qquad \cdot \wp(X_{n+m} = k | X_m = i) \\
&= \sum_k p_{ik}^{(n)} p_{kj} \\
&= p_{ij}^{(n+1)},
\end{aligned}$$

by the Chapman–Kolmogorov equation.

9.2. Continuous-Time Markov Chains

A family of integer-valued random variables, $\{X_t, t \geq 0\}$, is called a **Markov chain** if for all $n \geq 1$, and for all $0 \leq s_0 < \cdots < s_{n-1} < s < t$,

$$\wp(X_t = j | X_s = i, X_{s_{n-1}} = i_{n-1}, \ldots, X_{s_0} = i_0) \;=\; \wp(X_t = j | X_s = i).$$

In other words, given the sequence of values $i_0, \ldots, i_{n-1}, i$, the conditional probability of what X_t will be depends only on the condition $X_s = i$. The quantity $\wp(X_t = j | X_s = i)$ is called the **transition probability**.

***Example* 9.4.** Show that the Poisson process of rate λ is a Markov chain.

Solution. To begin, observe that

$$\wp(N_t = j | N_s = i, N_{s_{n-1}} = i_{n-1}, \ldots, N_{s_0} = i_0),$$

is equal to

$$\wp(N_t \Leftrightarrow i = j \Leftrightarrow i | N_s = i, N_{s_{n-1}} = i_{n-1}, \ldots, N_{s_0} = i_0).$$

By the substitution law, this is equal to

$$\wp(N_t \Leftrightarrow N_s = j \Leftrightarrow i | N_s = i, N_{s_{n-1}} = i_{n-1}, \ldots, N_{s_0} = i_0). \tag{9.8}$$

Since

$$(N_s, N_{s_{n-1}}, \ldots, N_{s_0}) \tag{9.9}$$

is a function of

$$(N_s \Leftrightarrow N_{s_{n-1}}, \ldots, N_{s_1} \Leftrightarrow N_{s_0}, N_{s_0} \Leftrightarrow N_0),$$

and since this is independent of $N_t \Leftrightarrow N_s$ by the independent increments property of the Poisson process, it follows that (9.9) and $N_t \Leftrightarrow N_s$ are also independent. Thus, (9.8) is equal to $\wp(N_t \Leftrightarrow N_s = j \Leftrightarrow i)$, which depends on i but not on $i_{n-1}, \ldots, i_0$. It then follows that

$$\wp(N_t = j | N_s = i, N_{s_{n-1}} = i_{n-1}, \ldots, N_{s_0} = i_0) \;=\; \wp(N_t = j | N_s = i),$$

and we see that the Poisson process is a Markov chain.

As shown in the above example,

$$\wp(N_t = j | N_s = i) \;=\; \wp(N_t \Leftrightarrow N_s = j \Leftrightarrow i) \;=\; \frac{[\lambda(t \Leftrightarrow s)]^{j-i} e^{-\lambda(t-s)}}{(j \Leftrightarrow i)!}$$

depends on t and s only through $t \Leftrightarrow s$. In general, if a Markov chain has the property that the transition probability $\wp(X_t = j | X_s = i)$ depends on t and s only through $t \Leftrightarrow s$, we say that the chain is **time-homogeneous** or that it has **stationary transition probabilities**. In this case, if we put

$$p_{ij}(t) \;:=\; \wp(X_t = j | X_0 = i),$$

then $\wp(X_t = j | X_s = i) = p_{ij}(t \Leftrightarrow s)$. Note that $p_{ij}(0) = \delta_{ij}$, the Kronecker delta.

In the remainder of the chapter, we assume that X_t is a time-homogeneous Markov chain with transition probability function $p_{ij}(t)$. For such a chain, we can derive the continuous-time **Chapman–Kolmogorov equation**,

$$p_{ij}(t+s) \;=\; \sum_k p_{ik}(t) p_{kj}(s).$$

To derive this, we first use the law of total conditional probability to write

$$\begin{aligned} p_{ij}(t+s) &= \wp(X_{t+s} = j | X_0 = i) \\ &= \sum_k \wp(X_{t+s} = j | X_t = k, X_0 = i) \wp(X_t = k | X_0 = i). \end{aligned}$$

Now use the Markov property and time homogeneity to obtain

$$\begin{aligned} p_{ij}(t+s) &= \sum_k \wp(X_{t+s}=j|X_t=k)\wp(X_t=k|X_0=i) \\ &= \sum_k p_{kj}(s)p_{ik}(t). \end{aligned}$$

The reader may wonder why the derivation of the continuous-time Chapman–Kolmogorov equation is so much simpler than the derivation of the discrete-time version. The reason is that in discrete time, the Markov property and time homogeneity are defined in a one-step manner. Hence, induction arguments are first needed to *derive* the discrete-time analogs of the continuous-time *definitions*!

Kolmogorov's Differential Equations

In the remainder of the chapter, we assume that for small $\Delta t > 0$,

$$p_{ij}(\Delta t) \approx g_{ij}\Delta t, \quad i \neq j, \quad \text{and} \quad p_{ii}(\Delta t) \approx 1 + g_{ii}\Delta t.$$

These approximations tell us the conditional probability of being in state j at time Δt in the very near future given that we are in state i at time zero. These assumptions are more precisely written as

$$\lim_{\Delta t \downarrow 0} \frac{p_{ij}(\Delta t)}{\Delta t} = g_{ij} \quad \text{and} \quad \lim_{\Delta t \downarrow 0} \frac{p_{ii}(\Delta t) \Leftrightarrow 1}{\Delta t} = g_{ii}. \tag{9.10}$$

Note that $g_{ij} \geq 0$, while $g_{ii} \leq 0$.

Using the Chapman–Kolmogorov equation, write

$$\begin{aligned} p_{ij}(t+\Delta t) &= \sum_k p_{ik}(t)p_{kj}(\Delta t) \\ &= p_{ij}(t)p_{jj}(\Delta t) + \sum_{k \neq j} p_{ik}(t)p_{kj}(\Delta t). \end{aligned}$$

Now subtract $p_{ij}(t)$ from both sides to get

$$p_{ij}(t+\Delta t) \Leftrightarrow p_{ij}(t) = p_{ij}(t)[p_{jj}(\Delta t) \Leftrightarrow 1] + \sum_{k \neq j} p_{ik}(t)p_{kj}(\Delta t). \tag{9.11}$$

Dividing by Δt and applying the limit assumptions (9.10), we obtain

$$p'_{ij}(t) = p_{ij}(t)g_{jj} + \sum_{k \neq j} p_{ik}(t)g_{kj}.$$

This is **Kolmogorov's forward differential equation**, which can be written more compactly as

$$p'_{ij}(t) = \sum_k p_{ik}(t)g_{kj}. \tag{9.12}$$

To derive the backward equation, observe that since $p_{ij}(t+\Delta t) = p_{ij}(\Delta t+t)$, we can write

$$\begin{aligned} p_{ij}(t+\Delta t) &= \sum_k p_{ik}(\Delta t)p_{kj}(t) \\ &= p_{ii}(\Delta t)p_{ij}(t) + \sum_{k\neq i} p_{ik}(\Delta t)p_{kj}(t). \end{aligned}$$

Now subtract $p_{ij}(t)$ from both sides to get

$$p_{ij}(t+\Delta t) \Leftrightarrow p_{ij}(t) = [p_{ii}(\Delta t) \Leftrightarrow 1]p_{ij}(t) + \sum_{k\neq i} p_{ik}(\Delta t)p_{kj}(t).$$

Dividing by Δt and applying the limit assumptions (9.10), we obtain

$$p'_{ij}(t) = g_{ii}p_{ij}(t) + \sum_{k\neq i} g_{ik}p_{kj}(t).$$

This is **Kolmogorov's backward differential equation**, which can be written more compactly as

$$p'_{ij}(t) = \sum_k g_{ik}p_{kj}(t). \tag{9.13}$$

The derivations of both the forward and backward differential equations require taking a limit in Δt inside the sum over k. For example, in deriving the backward equation, we tacitly assumed that

$$\lim_{\Delta t\downarrow 0} \sum_{k\neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) = \sum_{k\neq i} \lim_{\Delta t\downarrow 0} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t). \tag{9.14}$$

If the state space of the chain is finite, the above sum is finite and there is no problem. Otherwise, additional technical assumptions are required to justify this step. We now show that a sufficient assumption for deriving the backward equation is that

$$\sum_{j\neq i} g_{ij} = \Leftrightarrow g_{ii} < \infty.$$

A chain that satisfies this condition is said to be **conservative**. For any finite N, observe that

$$\sum_{k\neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \geq \sum_{\substack{|k|\leq N \\ k\neq i}} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t).$$

Since the right-hand side is a finite sum,

$$\lim_{\Delta t\downarrow 0} \sum_{k\neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \geq \sum_{\substack{|k|\leq N \\ k\neq i}} g_{ik}p_{kj}(t).$$

Letting $N \to \infty$ shows that

$$\lim_{\Delta t \downarrow 0} \sum_{k \neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \;\geq\; \sum_{k \neq i} g_{ik} p_{kj}(t). \tag{9.15}$$

To get an upper bound on the limit, take $N \geq i$ and write

$$\sum_{k \neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \;=\; \sum_{\substack{|k| \leq N \\ k \neq i}} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) + \sum_{|k| > N} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t).$$

Since $p_{kj}(t) \leq 1$,

$$\begin{aligned}
\sum_{k \neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) &\leq \sum_{\substack{|k| \leq N \\ k \neq i}} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) + \sum_{|k| > N} \frac{p_{ik}(\Delta t)}{\Delta t} \\
&= \sum_{\substack{|k| \leq N \\ k \neq i}} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) + \frac{1}{\Delta t} \left(1 \Leftrightarrow \sum_{|k| \leq N} p_{ik}(\Delta t) \right) \\
&= \sum_{\substack{|k| \leq N \\ k \neq i}} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) + \frac{1 \Leftrightarrow p_{ii}(\Delta t)}{\Delta t} \Leftrightarrow \sum_{\substack{|k| \leq N \\ k \neq i}} \frac{p_{ik}(\Delta t)}{\Delta t}.
\end{aligned}$$

Since these sums are finite,

$$\lim_{\Delta t \downarrow 0} \sum_{k \neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \;\leq\; \sum_{\substack{|k| \leq N \\ k \neq i}} g_{ik} p_{kj}(t) \Leftrightarrow g_{ii} \Leftrightarrow \sum_{\substack{|k| \leq N \\ k \neq i}} g_{ik}.$$

Letting $N \to \infty$ shows that

$$\lim_{\Delta t \downarrow 0} \sum_{k \neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \;\leq\; \sum_{k \neq i} g_{ik} p_{kj}(t) \Leftrightarrow g_{ii} \Leftrightarrow \sum_{k \neq i} g_{ik}.$$

If the chain is conservative, this simplifies to

$$\lim_{\Delta t \downarrow 0} \sum_{k \neq i} \frac{p_{ik}(\Delta t)}{\Delta t} p_{kj}(t) \;\leq\; \sum_{k \neq i} g_{ik} p_{kj}(t).$$

Combining this with (9.15) yields (9.14), thus justifying the backward equation.

Readers familiar with linear system theory may find it insightful to write the forward and backward equations in matrix form. Let $P(t)$ denote the matrix whose $i\,j$ entry is $p_{ij}(t)$, and let G denote the matrix whose $i\,j$ entry is g_{ij} (G is called the **generator matrix**, or **rate matrix**). Then the forward equation (9.12) becomes

$$P'(t) \;=\; P(t)G,$$

and the backward equation (9.13) becomes

$$P'(t) \;=\; GP(t),$$

The initial condition in both cases is $P(0) = I$. Under suitable assumptions, the solution of both equations is given by the **matrix exponential**,

$$P(t) \;=\; e^{Gt} \;:=\; \sum_{n=0}^{\infty} \frac{(Gt)^n}{n!}.$$

When the state space is finite, G is a finite-dimensional matrix, and the theory is straightforward. Otherwise, more careful analysis is required.

Stationary Distributions

Let us suppose that $p_{ij}(t) \to \pi_j$ as $t \to \infty$, independently of the initial state i. Then for large t, $p_{ij}(t)$ is approximately constant, and we therefore hope that its derivative is converging to zero. Assuming this to be true, the limiting form of the forward equation (9.12) is

$$0 \;=\; \sum_k \pi_k g_{kj}.$$

Combining this with the normalization condition $\sum_k \pi_k = 1$ allows us to solve for π_k much as in the discrete case.

9.3. Problems

Problems §9.1: Introduction to Discrete-Time Markov Chains

1. Let $X_0, Z_1, Z_2, \ldots$ be a sequence of independent discrete random variables. Put

   $$X_n \;=\; g(X_{n-1}, Z_n), \quad n = 1, 2, \ldots.$$

   Show that X_n is a Markov chain. For example, if $X_n = \max(0, X_{n-1} + Z_n)$, where X_0 and the Z_n are as in Example 9.1, then X_n is a random walk restricted to the nonnegative integers.

2. Let X_n be a time-homogeneous Markov chain with transition probabilities p_{ij}. Put $\nu_i := \wp(X_0 = i)$. Express

   $$\wp(X_0 = i, X_1 = j, X_2 = k, X_3 = l)$$

   in terms of ν_i and entries from the transition probability matrix.

3. Find the stationary distribution of the Markov chain in Figure 9.2.

4. Draw the state transition diagram and find the stationary distribution of the Markov chain whose transition matrix is

   $$P \;=\; \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 0 & 3/4 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

   Answer: $\pi_0 = 5/12$, $\pi_1 = 1/3$, $\pi_2 = 1/4$.

5. Draw the state transition diagram and find the stationary distribution of the Markov chain whose transition matrix is

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 3/4 & 0 \\ 1/4 & 3/4 & 0 \end{bmatrix}.$$

Answer: $\pi_0 = 1/5$, $\pi_1 = 7/10$, $\pi_2 = 1/10$.

6. Draw the state transition diagram and find the stationary distribution of the Markov chain whose transition matrix is

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 9/10 & 0 & 1/10 & 0 \\ 0 & 1/10 & 0 & 9/10 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}.$$

Answer: $\pi_0 = 9/28$, $\pi_1 = 5/28$, $\pi_2 = 5/28$, $\pi_3 = 9/28$.

7. Find the stationary distribution of the queuing system with finite buffer of size N, whose state transition diagram is shown in Figure 9.7.

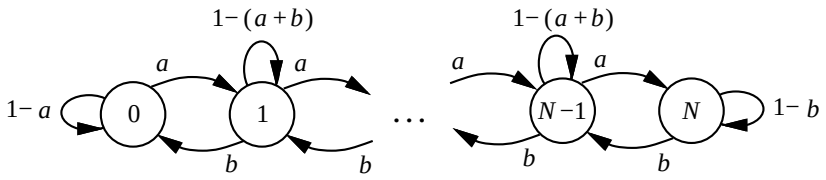

Figure 9.7. State transition diagram for a queue with a finite buffer.

8. Show that if $\nu_i := \wp(X_0 = i)$, then

$$\wp(X_n = j) = \sum_i \nu_i p_{ij}^{(n)}.$$

If $\{\pi_j\}$ is the stationary distribution, and if $\nu_i = \pi_i$, show that $\wp(X_n = j) = \pi_j$.

Problems §9.2: Continuous-Time Markov Chains

9. The general continuous-time random walk is defined by

$$g_{ij} = \begin{cases} \mu_i, & j = i \Leftrightarrow 1, \\ \Leftrightarrow(\lambda_i + \mu_i), & j = i, \\ \lambda_i, & j = i+1, \\ 0, & \text{otherwise.} \end{cases}$$

Write out the forward and backward equations. Is the chain conservative?

10. The continuous-time queue with infinite buffer can be obtained by modifying the general random walk in the preceding problem to include a barrier at the origin. Put

$$g_{0j} \;=\; \begin{cases} \Leftrightarrow\lambda_0, & j=0, \\ \lambda_0, & j=1, \\ 0, & \text{otherwise.} \end{cases}$$

Find the stationary distribution assuming

$$\sum_{j=1}^{\infty}\Big(\frac{\lambda_0\cdots\lambda_{j-1}}{\mu_1\cdots\mu_j}\Big) \;<\; \infty.$$

If $\lambda_i=\lambda$ and $\mu_i=\mu$ for all i, simplify the above condition to one involving only the relative values of λ and μ.

11. Modify Problem 10 to include a barrier at some finite N. Find the stationary distribution.

12. For the chain in Problem 10, let

$$\lambda_j = j\lambda+\alpha \quad\text{and}\quad \mu_j = j\mu,$$

where λ, α, and μ are positive. Put $m_i(t) := \mathsf{E}[X_t|X_0=i]$. Derive a differential equation for $m_i(t)$ and solve it. Treat the cases $\lambda=\mu$ and $\lambda\neq\mu$ separately. *Hint:* Use the forward equation (9.12).

13. For the chain in Problem 10, let $\mu_i=0$ and $\lambda_i=\lambda$. Write down and solve the forward equation (9.12). *Hint:* Equation (8.4).

14. Let T denote the first time a chain leaves state i,

$$T \;:=\; \min\{t\ge 0 : X_t\neq i\}.$$

Show that given $X_0=i$, T is conditionally $\exp(\Leftrightarrow g_{ii})$. In other words, the time the chain spends in state i, known as the **sojourn time**, has an exponential density with parameter $\Leftrightarrow g_{ii}$. *Hints:* By Problem 43 in Chapter 4, it suffices to prove that

$$\wp(T>t+\Delta t|T>t, X_0=i) \;=\; \wp(T>\Delta t|X_0=i).$$

To derive this equation, use the fact that if X_t is right-continuous,

$$T>t \quad\text{if and only if}\quad X_s=i \text{ for } 0\le s\le t.$$

Use the Markov property in the form

$$\begin{aligned}\wp(X_s=i, t\le s\le t+\Delta t|X_s=i, 0\le s\le t)& \\ = \;\; \wp(X_s=i, t\le s\le t+\Delta t|X_t=i),&\end{aligned}$$

and use time homogeneity in the form

$$\wp(X_s = i, t \le s \le t + \Delta t | X_t = i) \;=\; \wp(X_s = i, 0 \le s \le \Delta t | X_0 = i).$$

To identify the parameter of the exponential density, you may use the approximation $\wp(X_s = i, 0 \le s \le \Delta t | X_0 = i) \approx p_{ii}(\Delta t)$.

15. The notion a Markov chain can be generalized to include random variables that are not necessarily discrete. We say that X_t is a continuous-time **Markov process** if

$$\wp(X_t \in B | X_s = x, X_{s_{n-1}} = x_{n-1}, \ldots, X_{s_0} = x_0) \;=\; \wp(X_t \in B | X_s = x).$$

Such a process is time homogeneous if $\wp(X_t \in B | X_s = x)$ depends on t and s only through $t \Leftrightarrow s$. Show that the Wiener process is a Markov process that is time homogeneous. *Hint:* It is enough to look at conditional cdfs; i.e., show that

$$\wp(X_t \le y | X_s = x, X_{s_{n-1}} = x_{n-1}, \ldots, X_{s_0} = x_0) \;=\; \wp(X_t \le y | X_s = x).$$

16. Let X_t be a time-homogeneous Markov process as defined in the previous problem. Put

$$P_t(x, B) \;:=\; \wp(X_t \in B | X_0 = x),$$

and assume that there is a corresponding conditional density, denoted by $f_t(x, y)$, such that

$$P_t(x, B) \;=\; \int_B f_t(x, y)\, dy.$$

Derive the Chapman–Kolmogorov equation for conditional densities,

$$f_{t+s}(x, y) \;=\; \int_{-\infty}^{\infty} f_s(x, z) f_t(z, y)\, dz\,.$$

Hint: It suffices to show that

$$P_{t+s}(x, B) \;=\; \int_{-\infty}^{\infty} f_s(x, z) P_t(z, B)\, dz\,.$$

To derive this, you may assume that a law of total conditional probability holds for random variables with appropriate conditional densities.

CHAPTER 10

Mean Convergence and Applications

As mentioned at the beginning of Chapter 1, limit theorems have been the foundation of the success of Kolmogorov's axiomatic theory of probability. In this chapter and the next, we focus on different notions of convergence and their implications.

Section 10.1 introduces the notion of convergence in mean of order p. Section 10.2 introduces the normed L^p spaces. Norms provide a compact notation for establishing results about convergence in mean of order p. We also point out that the L^p spaces are complete. Completeness is used to show that convolution sums like

$$\sum_{k=0}^{\infty} h_k X_{n-k}$$

are well defined. This is an important result because sums like this represent the response of a causal, linear, time-invariant system to a random input X_k. Section 10.3 uses completeness to develop the Wiener integral. Section 10.4 introduces the notion of projections. The L^2 setting allows us to introduce a general orthogonality principle that unifies results from earlier chapters on the Wiener filter, linear estimators of random vectors, and minimum mean squared error estimation. The completeness of L^2 is also used to prove the projection theorem. In Section 10.5, the projection theorem is used to to establish the existence conditional expectation for random variables that may not be discrete or jointly continuous. In Section 10.6, completeness is used to establish the spectral representation of wide-sense stationary random sequences.

10.1. Convergence in Mean of Order p

We say that X_n **converges in mean of order p** to X if

$$\lim_{n\to\infty} \mathsf{E}[|X_n - X|^p] \;=\; 0,$$

where $1 \le p < \infty$. Mostly we focus on the cases $p = 1$ and $p = 2$. The case $p = 1$ is called **convergence in mean** or **mean convergence**. The case $p = 2$ is called **mean square convergence** or **quadratic mean convergence**.

***Example* 10.1.** Let $X_n \sim N(0, 1/n^2)$. Show that $\sqrt{n}X_n$ converges in mean square to zero.

Solution. Write

$$\mathsf{E}[|\sqrt{n}X_n|^2] \;=\; n\mathsf{E}[X_n^2] \;=\; n\frac{1}{n^2} \;=\; \frac{1}{n} \;\to\; 0.$$

In the following example, X_n converges in mean square to zero, but not in mean of order 4.

***Example* 10.2.** Let X_n have density

$$f_n(x) \;=\; g_n(x)(1 \Leftrightarrow 1/n^3) + h_n(x)/n^3,$$

where $g_n \sim N(0, 1/n^2)$ and $h_n \sim N(n, 1)$. Show that X_n converges to zero in mean square, but not in mean of order 4.

Solution. For convergence in mean square, write

$$\mathsf{E}[|X_n|^2] \;=\; \frac{1}{n^2}(1 \Leftrightarrow 1/n^3) + (1 + n^2)/n^3 \;\to\; 0.$$

However, using Problem 28 in Chapter 3, we have

$$\mathsf{E}[|X_n|^4] \;=\; \frac{3}{n^4}(1 \Leftrightarrow 1/n^3) + (n^4 + 6n^2 + 3)/n^3 \;\to\; \infty.$$

The preceding example raises the question of whether X_n might converge in mean of order 4 to something other than zero. However, by Problem 8 at the end of the chapter, if X_n converged in mean of order 4 to some X, then it would also converge in mean square to X. Hence, the only possible limit for X_n in mean of order 4 is zero, and as we saw, X_n does not converge in mean of order 4 to zero.

***Example* 10.3.** Let $X_1, X_2, \ldots$ be uncorrelated random variables with common mean m and common variance σ^2. Show that the sample mean

$$M_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in mean square to m. We call this the **mean-square law of large numbers** for uncorrelated random variables.

Solution. Since

$$M_n \Leftrightarrow m \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i \Leftrightarrow m),$$

we can write

$$\begin{aligned}
\mathsf{E}[|M_n \Leftrightarrow m|^2] &= \frac{1}{n^2}\mathsf{E}\left[\left(\sum_{i=1}^{n}(X_i \Leftrightarrow m)\right)\left(\sum_{j=1}^{n}(X_j \Leftrightarrow m)\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathsf{E}[(X_i \Leftrightarrow m)(X_j \Leftrightarrow m)]. \qquad (10.1)
\end{aligned}$$

Since X_i and X_j are uncorrelated, the preceding expectations are zero when $i \neq j$. Hence,

$$\mathsf{E}[|M_n \Leftrightarrow m|^2] \;=\; \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{E}[(X_i \Leftrightarrow m)^2] \;=\; \frac{n\sigma^2}{n^2} \;=\; \frac{\sigma^2}{n},$$

which goes to zero as $n \to \infty$.

***Example* 10.4.** Here is a generalization of the preceding example. Let $X_1, X_2, \ldots$ be wide-sense stationary; i.e., the X_i have common mean $m = \mathsf{E}[X_i]$, and the covariance $\mathsf{E}[(X_i \Leftrightarrow m)(X_j \Leftrightarrow m)]$ depends only on the difference $i \Leftrightarrow j$. Put

$$R(i) \;:=\; \mathsf{E}[(X_{j+i} \Leftrightarrow m)(X_j \Leftrightarrow m)].$$

Show that

$$M_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in mean square to m if and only if

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} R(k) \;=\; 0. \tag{10.2}$$

We call this result the **mean-square law of large numbers** for wide-sense stationary sequences or the **mean-square ergodic theorem for wide-sense stationary sequences** Note that a sufficient condition for (10.2) to hold is that $\lim_{k\to\infty} R(k) = 0$ (Problem 3).

Solution. We show that (10.2) implies M_n converges in mean square to m. The converse is left to the reader in Problem 4. From (10.1), we see that

$$\begin{aligned} n^2\,\mathsf{E}[|M_n \Leftrightarrow m|^2] &= \sum_{i=1}^{n}\sum_{j=1}^{n} R(i \Leftrightarrow j) \\ &= \sum_{i=j} R(0) + 2\sum_{j<i} R(i \Leftrightarrow j) \\ &= nR(0) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} R(i \Leftrightarrow j) \\ &= nR(0) + 2\sum_{i=2}^{n}\sum_{k=1}^{i-1} R(k). \end{aligned}$$

On account of (10.2), given $\varepsilon > 0$, there is an N such that for all $i \geq N$,

$$\left|\frac{1}{i \Leftrightarrow 1}\sum_{k=1}^{i-1} R(k)\right| \;<\; \varepsilon.$$

For $n \geq N$, the above double sum can be written as

$$\sum_{i=2}^{n}\sum_{k=1}^{i-1} R(k) \;=\; \sum_{i=2}^{N-1}\sum_{k=1}^{i-1} R(k) + \sum_{i=N}^{n}(i \Leftrightarrow 1)\left[\frac{1}{i \Leftrightarrow 1}\sum_{k=1}^{i-1} R(k)\right].$$

The magnitude of the right-most double sum is upper bounded by

$$\begin{aligned}\left|\sum_{i=N}^{n}(i \Leftrightarrow 1)\left[\frac{1}{i \Leftrightarrow 1}\sum_{k=1}^{i-1} R(k)\right]\right| &< \varepsilon \sum_{i=N}^{n}(i \Leftrightarrow 1) \\ &< \varepsilon \sum_{i=1}^{n}(i \Leftrightarrow 1) \\ &= \frac{\varepsilon n(n \Leftrightarrow 1)}{2} \\ &< \frac{\varepsilon n^2}{2}.\end{aligned}$$

It now follows that $\lim_{n\to\infty} \mathsf{E}[|M_n \Leftrightarrow m|^2]$ can be no larger than ε. Since ε is arbitrary, the limit must be zero.

***Example* 10.5.** Let W_t be a Wiener process with $\mathsf{E}[W_t^2] = \sigma^2 t$. Show that W_t/t converges in mean square to zero as $t \to \infty$.

Solution. Write

$$\mathsf{E}\left[\left|\frac{W_t}{t}\right|^2\right] \;=\; \frac{\sigma^2 t}{t^2} \;=\; \frac{\sigma^2}{t} \;\to\; 0.$$

***Example* 10.6.** Let X be a nonnegative random variable with finite mean; i.e., $\mathsf{E}[X] < \infty$. Put

$$X_n \;:=\; \min(X, n) \;=\; \begin{cases} X, & X \leq n, \\ n, & X > n. \end{cases}$$

The idea here is that X_n is a bounded random variable that can be used to approximate X. Show that X_n converges in mean to X.

Solution. Since $X \geq X_n$, $\mathsf{E}[|X_n \Leftrightarrow X|] = \mathsf{E}[X \Leftrightarrow X_n]$. Since $X \Leftrightarrow X_n$ is nonnegative, we can write

$$\mathsf{E}[X \Leftrightarrow X_n] \;=\; \int_0^\infty \wp(X \Leftrightarrow X_n > t)\, dt,$$

where we have appealed to (4.2) in Section 4.2. Next, for $t \geq 0$, a little thought shows that

$$\{X \Leftrightarrow X_n > t\} \;=\; \{X > t + n\}.$$

Hence,

$$\mathsf{E}[X \Leftrightarrow X_n] \;=\; \int_0^\infty \wp(X > t + n)\, dt \;=\; \int_n^\infty \wp(X > \theta)\, d\theta,$$

which goes to zero as $n \to \infty$ on account of the fact that

$$\infty \;>\; \mathsf{E}[X] \;=\; \int_0^\infty \wp(X > \theta)\, d\theta.$$

10.2. Normed Vector Spaces of Random Variables

We denote by L^p the set of all random variables X with the property that $\mathsf{E}[|X|^p] < \infty$. We claim that L^p is a vector space. To prove this, we need to show that if $\mathsf{E}[|X|^p] < \infty$ and $\mathsf{E}[|Y|^p] < \infty$, then $\mathsf{E}[|aX + bY|^p] < \infty$ for all scalars a and b. To begin, recall that the **triangle inequality** applied to numbers x and y says that

$$|x + y| \;\le\; |x| + |y|.$$

If $|y| \le |x|$, then

$$|x + y| \;\le\; 2|x|,$$

and so

$$|x + y|^p \;\le\; 2^p |x|^p.$$

A looser bound that has the advantage of being symmetric is

$$|x + y|^p \;\le\; 2^p \left(|x|^p + |y|^p\right).$$

It is easy to see that this bound also holds if $|y| > |x|$. We can now write

$$\begin{aligned} \mathsf{E}[|aX + bY|^p] &\le \mathsf{E}[2^p \left(|aX|^p + |bY|^p\right)] \\ &= 2^p \left(|a|^p \mathsf{E}[|X|^p] + |b|^p \mathsf{E}[|Y|^p]\right). \end{aligned}$$

Hence, if $\mathsf{E}[|X|^p]$ and $\mathsf{E}[|Y|^p]$ are both finite, then so is $\mathsf{E}[|aX + bY|^p]$.

For $X \in L^p$, we put

$$\|X\|_p \;:=\; \mathsf{E}[|X|^p]^{1/p}.$$

We claim that $\|\cdot\|_p$ is a **norm** on L^p, by which we mean the following three properties hold:

(*i*) $\|X\|_p \ge 0$, and $\|X\|_p = 0$ if and only if X is the zero random variable.

(*ii*) For scalars a, $\|aX\|_p = |a|\, \|X\|_p$.

(*iii*) For $X, Y \in L^p$,

$$\|X + Y\|_p \;\le\; \|X\|_p + \|Y\|_p.$$

As in the numerical case, this is also known as the **triangle inequality**.

The first two properties are obvious, while the third one is known as **Minkowski's inequality**, which is derived in Problem 9.

Observe now that X_n converges in mean of order p to X if and only if

$$\lim_{n\to\infty} \|X_n \ominus X\|_p \;=\; 0.$$

Hence, the three norm properties above can be used to derive results about convergence in mean of order p, as will be seen in the problems.

Recall that a sequence of real numbers x_n is **Cauchy** if for every $\varepsilon > 0$, for all sufficiently large n and m, $|x_n \ominus x_m| < \varepsilon$. A basic fact that can be proved about the set of real numbers is that it is **complete**; i.e., given any Cauchy sequence of real numbers x_n, there is a real number x such that x_n converges to x [38, p. 53, Theorem 3.11]. Similarly, a sequence of random variables $X_n \in L^p$ is said to be **Cauchy** if for every $\varepsilon > 0$, for all sufficiently large n and m,

$$\|X_n \ominus X_m\|_p \;<\; \varepsilon.$$

It can be shown that the L^p spaces are complete; i.e., if X_n is a Cauchy sequence of L^p random variables, then there exists an L^p random variable X such that X_n converges in mean of order p to X. This is known as the **Riesz–Fischer Theorem** [37, p. 244]. A normed vector space that is complete is called a **Banach space**.

Of special interest is the case $p = 2$ because the norm $\|\cdot\|_2$ can be expressed in terms of the **inner product***

$$\langle X, Y\rangle \;:=\; \mathsf{E}[XY], \quad X, Y \in L^2.$$

It is easily seen that

$$\langle X, X\rangle^{1/2} \;=\; \|X\|_2.$$

Because the norm $\|\cdot\|_2$ can be obtained using the inner product, L^2 is called an **inner-product space**. Since the L^p spaces are complete, L^2 in particular is a complete inner-product space. A complete inner-product space is called a **Hilbert space**.

The space L^2 has several important properties. First, for fixed Y, it is easy to see that $\langle X, Y\rangle$ is linear in X. Second, the simple relationship between the norm and the inner product imply the **parallelogram law** (Problem 25),

$$\|X + Y\|_2^2 + \|X \ominus Y\|_2^2 \;=\; 2(\|X\|_2^2 + \|Y\|_2^2).$$

Third, there is the **Cauchy–Schwarz inequality** (Problem 1 in Chapter 6),

$$\big|\langle X, Y\rangle\big| \;\le\; \|X\|_2 \|Y\|_2.$$

*For complex-valued random variables (defined in Section 7.5), we put $\langle X, Y\rangle := \mathsf{E}[XY^*]$.

***Example* 10.7.** Show that

$$\sum_{k=1}^{\infty} h_k X_k$$

is well defined as an element of L^2 assuming that

$$\sum_{k=1}^{\infty} |h_k| \;<\; \infty \quad \text{and} \quad \mathsf{E}[|X_k|^2] \;\le\; B, \quad \text{for all } k,$$

where B is a finite constant.

Solution. Consider the partial sums,

$$Y_n \;:=\; \sum_{k=1}^{n} h_k X_k.$$

Each Y_n is an element of L^2, which is complete. If we can show that Y_n is a Cauchy sequence, then there will exist a $Y \in L^2$ with $\|Y_n \Leftrightarrow Y\|_2 \to 0$. Thus, the infinite-sum expression $\sum_{k=1}^{\infty} h_k X_k$ is understood to be shorthand for "the mean-square limit of $\sum_{k=1}^{n} h_k X_k$ as $n \to \infty$." Next, for $n > m$, write

$$Y_n \Leftrightarrow Y_m \;=\; \sum_{k=m+1}^{n} h_k X_k.$$

Then

$$\begin{aligned}
\|Y_n \Leftrightarrow Y_m\|_2^2 &= \langle Y_n \Leftrightarrow Y_m, Y_n \Leftrightarrow Y_m \rangle \\
&= \left\langle \sum_{k=m+1}^{n} h_k X_k, \sum_{l=m+1}^{n} h_l X_l \right\rangle \\
&\le \sum_{k=m+1}^{n} \sum_{l=m+1}^{n} |h_k|\,|h_l|\,\big|\langle X_k, X_l \rangle\big| \\
&\le \sum_{k=m+1}^{n} \sum_{l=m+1}^{n} |h_k|\,|h_l|\,\|X_k\|_2\,\|X_l\|_2,
\end{aligned}$$

by the Cauchy–Schwarz inequality. Next, since $\|X_k\|_2 = \mathsf{E}[|X_k|^2]^{1/2} \le \sqrt{B}$,

$$\|Y_n \Leftrightarrow Y_m\|_2^2 \;\le\; B\left(\sum_{k=m+1}^{n} |h_k| \right)^2.$$

Since $\sum_{k=1}^{\infty} |h_k| < \infty$ implies[1]

$$\sum_{k=m+1}^{n} |h_k| \;\to\; 0 \quad \text{as } n \text{ and } m \to \infty \text{ with } n > m,$$

it follows that Y_n is Cauchy.

Under the conditions of the previous example, since $\|Y_n \Leftrightarrow Y\|_2 \to 0$, we can write, for any $Z \in L^2$,

$$\begin{aligned} \big|\langle Y_n, Z\rangle \Leftrightarrow \langle Y, Z\rangle\big| &= \big|\langle Y_n \Leftrightarrow Y, Z\rangle\big| \\ &\le \|Y_n \Leftrightarrow Y\|_2 \, \|Z\|_2 \;\to\; 0. \end{aligned}$$

In other words,

$$\lim_{n\to\infty} \langle Y_n, Z\rangle \;=\; \langle Y, Z\rangle.$$

Taking $Z = 1$ and writing the inner products as expectations yields

$$\lim_{n\to\infty} \mathsf{E}[Y_n] \;=\; \mathsf{E}[Y],$$

or, using the definitions of Y_n and Y,

$$\lim_{n\to\infty} \mathsf{E}\left[\sum_{k=1}^{n} h_k X_k\right] \;=\; \mathsf{E}\left[\sum_{k=1}^{\infty} h_k X_k\right].$$

Since the sum on the left is finite, we can bring the expectation inside and write

$$\lim_{n\to\infty} \sum_{k=1}^{n} h_k \mathsf{E}[X_k] \;=\; \mathsf{E}\left[\sum_{k=1}^{\infty} h_k X_k\right].$$

In other words,

$$\sum_{k=1}^{\infty} h_k \mathsf{E}[X_k] \;=\; \mathsf{E}\left[\sum_{k=1}^{\infty} h_k X_k\right].$$

The foregoing example and discussion have the following application. Consider a discrete-time, causal, stable, linear, time-invariant system with impulse response h_k. Now suppose that the random sequence X_k is applied to the input of this system. If $\mathsf{E}[|X_k|^2]$ is bounded as in the example, then[†] the output of the system at time n is

$$\sum_{k=0}^{\infty} h_k X_{n-k},$$

which is a well-defined element of L^2. Furthermore,

$$\mathsf{E}\left[\sum_{k=0}^{\infty} h_k X_{n-k}\right] \;=\; \sum_{k=0}^{\infty} h_k \mathsf{E}[X_{n-k}].$$

[†] The assumption in the example, $\sum_k |h_k| < \infty$, is equivalent to the assumption that the system is stable.

10.3. The Wiener Integral (Again)

In Section 8.3, we defined the Wiener integral

$$\int_0^\infty g(\tau)\, dW_\tau \;:=\; \sum_i g_i(W_{t_{i+1}} \Leftrightarrow W_{t_i}),$$

for piecewise constant g, say $g(\tau) = g_i$ for $t_i < t \le t_{i+1}$ for a finite number of intervals, and $g(\tau) = 0$ otherwise. In this case, since the integral is the sum of scaled, independent, zero mean, Gaussian increments of variance $\sigma^2(t_{i+1} \Leftrightarrow t_i)$,

$$\mathsf{E}\left[\left(\int_0^\infty g(\tau)\, dW_\tau\right)^2\right] \;=\; \sigma^2 \sum_i g_i^2(t_{i+1} \Leftrightarrow t_i) \;=\; \int_0^\infty g(\tau)^2\, d\tau.$$

We now define the Wiener integral for arbitrary g satisfying

$$\int_0^\infty g(\tau)^2\, d\tau \;<\; \infty. \tag{10.3}$$

To do this, we use the fact [13, p. 86, Prop. 3.4.2] that for g satisfying (10.3), there always exists a sequence of piecewise constant functions g_n converging to g in the mean-square sense

$$\lim_{n\to\infty} \int_0^\infty |g_n(\tau) \Leftrightarrow g(\tau)|^2\, d\tau \;=\; 0. \tag{10.4}$$

The set of g satisfying (10.3) is an inner product space with inner product $\langle g, h\rangle = \int_0^\infty g(\tau)h(\tau)\, d\tau$ and corresponding norm $\|g\| = \langle g, g\rangle^{1/2}$. Thus, (10.4) says that $\|g_n \Leftrightarrow g\| \to 0$. In particular, this implies g_n is Cauchy; i.e., $\|g_n \Leftrightarrow g_m\| \to 0$ as $n, m \to \infty$ (cf. Problem 11). Consider the random variables

$$Y_n \;:=\; \int_0^\infty g_n(\tau)\, dW_\tau.$$

Since each g_n is piecewise constant, Y_n is well defined and is Gaussian with zero mean and variance

$$\int_0^\infty g_n(\tau)^2\, d\tau.$$

Now observe that

$$\begin{aligned}
\|Y_n \Leftrightarrow Y_m\|_2^2 &= \mathsf{E}[|Y_n \Leftrightarrow Y_m|^2] \\
&= \mathsf{E}\left[\left|\int_0^\infty g_n(\tau)\, dW_\tau \Leftrightarrow \int_0^\infty g_m(\tau)\, dW_\tau\right|^2\right] \\
&= \mathsf{E}\left[\left|\int_0^\infty \big[g_n(\tau) \Leftrightarrow g_m(\tau)\big]\, dW_\tau\right|^2\right] \\
&= \int_0^\infty |g_n(\tau) \Leftrightarrow g_m(\tau)|^2\, d\tau,
\end{aligned}$$

since $g_n \ominus g_m$ is piecewise constant. Thus,

$$\|Y_n \ominus Y_m\|_2^2 \;=\; \|g_n \ominus g_m\|^2.$$

Since g_n is Cauchy, we see that Y_n is too. Since L^2 is complete, there exists a random variable $Y \in L^2$ with $\|Y_n \ominus Y\|_2 \to 0$. We denote this random variable by

$$\int_0^\infty g(\tau)\, dW_\tau\,,$$

and call it the Wiener integral of g.

10.4. Projections, Orthonality Principle, Projection Theorem

Let C be a subset of L^p. Given $X \in L^p$, suppose we want to approximate X by some $\widehat{X} \in C$. We call $\widehat{X}$ a **projection** of X onto C if $\widehat{X} \in C$ and if

$$\|X \ominus \widehat{X}\|_p \;\le\; \|X \ominus Y\|_p, \quad \text{for all } Y \in C.$$

Note that if $X \in C$, the we can take $\widehat{X} = X$.

***Example* 10.8.** Let C be the unit ball,

$$C \;:=\; \{Y \in L^p : \|Y\|_p \le 1\}.$$

For $X \notin C$, i.e., $\|X\|_p > 1$, show that

$$\widehat{X} \;=\; \frac{X}{\|X\|_p}.$$

Solution. First note that the proposed formula for $\widehat{X}$ satisfies $\|\widehat{X}\|_p = 1$ so that $\widehat{X} \in C$ as required. Now observe that

$$\begin{aligned}
\|X \ominus \widehat{X}\|_p &= \left\| X \ominus \frac{X}{\|X\|_p} \right\|_p \\
&= \left| 1 \ominus \frac{1}{\|X\|_p} \right| \|X\|_p \\
&= \|X\|_p \ominus 1.
\end{aligned}$$

Next, for any $Y \in C$,

$$\begin{aligned}
\|X \ominus Y\|_p &\ge \big|\|X\|_p \ominus \|Y\|_p\big|, \quad \text{by Problem 15,} \\
&= \|X\|_p \ominus \|Y\|_p \\
&\ge \|X\|_p \ominus 1 \\
&= \|X \ominus \widehat{X}\|_p.
\end{aligned}$$

Thus, no $Y \in C$ is closer to X than $\widehat{X}$.

Much more can be said about projections when $p = 2$ and when the set we are projecting onto is a subspace rather than an arbitrary subset.

We now present two fundamental results about projections onto subspaces of L^2. The first result is the **orthogonality principle**. Let M be a subspace of L^2. If $X \in L^2$, then $\widehat{X} \in M$ satisfies

$$\|X \Leftrightarrow \widehat{X}\|_2 \;\leq\; \|X \Leftrightarrow Y\|_2, \quad \text{for all } Y \in M, \tag{10.5}$$

if and only if

$$\langle X \Leftrightarrow \widehat{X}, Y \rangle \;=\; 0, \quad \text{for all } Y \in M. \tag{10.6}$$

Furthermore, if such an $\widehat{X} \in M$ exists, it is unique. Observe that there is no claim that an $\widehat{X} \in M$ exists that satisfies either (10.5) or (10.6). In practice, we try to find an $\widehat{X} \in M$ satisfying (10.6), since such an $\widehat{X}$ then automatically satisfies (10.5). This was the approach used to derive the Wiener filter in Section 6.5, where we implicitly took (for fixed t)

$$M \;=\; \big\{\widehat{V}_t : \widehat{V}_t \text{ is given by (6.7) and } \mathsf{E}[\widehat{V}_t^2] < \infty\big\}.$$

In Section 7.3, when we discussed linear estimation of random vectors, we implicitly took

$$M \;=\; \{AY + b : A \text{ is a matrix and } b \text{ is a column vector }\}.$$

When we discussed minimum mean squared error estimation, also in Section 7.3, we implicitly took

$$M \;=\; \big\{g(Y) : g \text{ is any function such that } \mathsf{E}[g(Y)^2] < \infty\big\}.$$

Thus, several estimation problems discussed in earlier chapters are seen to be special cases of finding the projection onto a suitable subspace of L^2. Each of these special cases had its version of the orthogonality principle, and so it should be no trouble for the reader to show that (10.6) implies (10.5). The converse is also true, as we now show. Suppose (10.5) holds, but for some $Y \in M$,

$$\langle X \Leftrightarrow \widehat{X}, Y \rangle \;=\; c \;\neq\; 0.$$

Because we can divide this equation by $\|Y\|_2$, there is no loss of generality in assuming $\|Y\|_2 = 1$. Now, since M is a subspace containing both $\widehat{X}$ and Y, $\widehat{X} + cY$ also belongs to M. We show that this new vector is strictly closer to X than $\widehat{X}$, contradicting (10.5). Write

$$\begin{aligned}
\|X \Leftrightarrow (\widehat{X} + cY)\|_2^2 &= \|(X \Leftrightarrow \widehat{X}) \Leftrightarrow cY\|_2^2 \\
&= \|X \Leftrightarrow \widehat{X}\|_2^2 \Leftrightarrow |c|^2 \Leftrightarrow |c|^2 + |c|^2 \\
&= \|X \Leftrightarrow \widehat{X}\|_2^2 \Leftrightarrow |c|^2 \\
&> \|X \Leftrightarrow \widehat{X}\|_2^2.
\end{aligned}$$

The second fundamental result to be presented is the **Projection Theorem**. Recall that the orthogonality principle does not guarantee the existence of an $\widehat{X} \in M$ satisfying (10.5). If we are not smart enough to solve (10.6), what can we do? This is where the projection theorem comes in. To state and prove this result, we need the concept of a **closed set**. We say that M is closed if whenever $X_n \in M$ and $\|X_n \Leftrightarrow X\|_2 \to 0$ for some $X \in L^2$, the limit X must actually be in M. In other words, a set is closed if it contains all the limits of all converging sequences from the set.

***Example* 10.9.** Show that the set of Wiener integrals

$$M \;:=\; \left\{ \int_0^\infty g(\tau)\, dW_\tau \;:\; \int_0^\infty g(\tau)^2\, d\tau < \infty \right\}$$

is closed.

Solution. A sequence X_n from M has the form

$$X_n \;=\; \int_0^\infty g_n(\tau)\, dW_\tau$$

for square-integrable g. Suppose X_n converges in mean square to some X. We must show that there exists a square-integrable function g for which

$$X \;=\; \int_0^\infty g(\tau)\, dW_\tau.$$

Since X_n converges, it is Cauchy. Now observe that

$$\begin{aligned}
\|X_n \Leftrightarrow X_m\|_2^2 &= \mathsf{E}[|X_n \Leftrightarrow X_m|^2] \\
&= \mathsf{E}\left[\left|\int_0^\infty g_n(\tau)\, dW_\tau \Leftrightarrow \int_0^\infty g_m(\tau)\, dW_\tau\right|^2\right] \\
&= \mathsf{E}\left[\left|\int_0^\infty \big[g_n(\tau) \Leftrightarrow g_m(\tau)\big]\, dW_\tau\right|^2\right] \\
&= \int_0^\infty |g_n(\tau) \Leftrightarrow g_m(\tau)|^2\, d\tau \\
&= \|g_n \Leftrightarrow g_m\|^2.
\end{aligned}$$

Thus, g_n is Cauchy. Since the set of square-integrable time functions is complete (the Riesz–Fischer Theorem again [37, p. 244]), there is a square-integrable g with $\|g_n \Leftrightarrow g\| \to 0$. For this g, write

$$\begin{aligned}
\left\|X_n \Leftrightarrow \int_0^\infty g(\tau)\, dW_\tau\right\|_2^2 &= \mathsf{E}\left[\left|\int_0^\infty g_n(\tau)\, dW_\tau \Leftrightarrow \int_0^\infty g(\tau)\, dW_\tau\right|^2\right] \\
&= \mathsf{E}\left[\left|\int_0^\infty \big[g_n(\tau) \Leftrightarrow g(\tau)\big]\, dW_\tau\right|^2\right] \\
&= \int_0^\infty |g_n(\tau) \Leftrightarrow g(\tau)|^2\, d\tau \\
&= \|g_n \Leftrightarrow g\|^2 \;\to\; 0.
\end{aligned}$$

Since mean-square limits are unique (Problem 14), $X = \int_0^\infty g(\tau)\, dW_\tau$.

Projection Theorem. *If M is a closed subspace of L^2, and $X \in L^2$, then there exists a unique $\widehat{X} \in M$ such that (10.5) holds.*

To prove this result, first put $h := \inf_{Y \in M} \|X \Leftrightarrow Y\|_2$. From the definition of the infimum, there is a sequence $Y_n \in M$ with $\|X \Leftrightarrow Y_n\|_2 \to h$. We will show that Y_n is a Cauchy sequence. Since L^2 is a Hilbert space, Y_n converges to some limit in L^2. Since M is closed, the limit, say $\widehat{X}$, must be in M.

To show Y_n is Cauchy, we proceed as follows. By the parallelogram law,

$$\begin{aligned} 2(\|X \Leftrightarrow Y_n\|_2^2 + \|X \Leftrightarrow Y_m\|_2^2) &= \|2X \Leftrightarrow (Y_n + Y_m)\|_2^2 + \|Y_m \Leftrightarrow Y_n\|_2^2 \\ &= 4\left\| X \Leftrightarrow \frac{Y_n + Y_m}{2} \right\|_2^2 + \|Y_m \Leftrightarrow Y_n\|_2^2. \end{aligned}$$

Note that the vector $(Y_n + Y_m)/2 \in M$ since M is a subspace. Hence,

$$2(\|X \Leftrightarrow Y_n\|_2^2 + \|X \Leftrightarrow Y_m\|_2^2) \;\geq\; 4h^2 + \|Y_m \Leftrightarrow Y_n\|_2^2.$$

Since $\|X \Leftrightarrow Y_n\|_2 \to h$, given $\varepsilon > 0$, there exists an N such that for all $n \geq N$, $\|X \Leftrightarrow Y_n\|_2 < h + \varepsilon$. Hence, for $m, n \geq N$,

$$\begin{aligned} \|Y_m \Leftrightarrow Y_n\|_2^2 &< 2((h+\varepsilon)^2 + (h+\varepsilon)^2) \Leftrightarrow 4h^2 \\ &= 4\varepsilon(2h + \varepsilon), \end{aligned}$$

and we see that Y_n is Cauchy.

Since L^2 is a Hilbert space, and since M is closed, $Y_n \to \widehat{X}$ for some $\widehat{X} \in M$. We now have to show that $\|X \Leftrightarrow \widehat{X}\|_2 \leq \|X \Leftrightarrow Y\|_2$ for all $Y \in M$. Write

$$\begin{aligned} \|X \Leftrightarrow \widehat{X}\|_2 &= \|X \Leftrightarrow Y_n + Y_n \Leftrightarrow \widehat{X}\|_2 \\ &\leq \|X \Leftrightarrow Y_n\|_2 + \|Y_n \Leftrightarrow \widehat{X}\|_2. \end{aligned}$$

Since $\|X \Leftrightarrow Y_n\|_2 \to h$ and since $\|Y_n \Leftrightarrow \widehat{X}\|_2 \to 0$,

$$\|X \Leftrightarrow \widehat{X}\|_2 \;\leq\; h.$$

Since $h \leq \|X \Leftrightarrow Y\|_2$ for all $Y \in M$, $\|X \Leftrightarrow \widehat{X}\|_2 \leq \|X \Leftrightarrow Y\|_2$ for all $Y \in M$.

The uniqueness of $\widehat{X}$ is left to Problem 22.

10.5. Conditional Expectation

In earlier chapters, we defined conditional expectation separately for discrete and jointly continuous random variables. We are now in a position to introduce a unified definition. The new definition is closely related to the orthogonality principle and the projection theorem. It also reduces to the earlier definitions in the discrete and jointly continuous cases.

We say that $\widehat{g}(Y)$ is the **conditional expectation of X given Y** if

$$\mathsf{E}[Xg(Y)] \;=\; \mathsf{E}[\widehat{g}(Y)g(Y)], \quad \text{for all bounded functions } g. \tag{10.7}$$

When X and Y are discrete or jointly continuous it is easy to check that $\widehat{g}(y) = \mathsf{E}[X|Y=y]$ solves this equation.[‡] However, the importance of this definition is that we can prove the existence and uniqueness of such a function $\widehat{g}$ even if X and Y are not discrete or jointly continuous, as long as $X \in L^1$. Recall that uniqueness was proved in Problem 18 in Chapter 7.

We first consider the case $X \in L^2$. Put

$$M \;:=\; \{g(Y) : \mathsf{E}[g(Y)^2] < \infty\}. \tag{10.8}$$

It is a consequence of the Riesz–Fischer Theorem [37, p. 244] that M is closed. By the projection theorem combined with the orthogonality principle, there exists a $\widehat{g}(Y) \in M$ such that

$$\langle X \Leftrightarrow \widehat{g}(Y), g(Y) \rangle \;=\; 0, \quad \text{for all } g(Y) \in M.$$

Since the above inner product is defined as an expectation, it is equivalent to

$$\mathsf{E}[Xg(Y)] \;=\; \mathsf{E}[\widehat{g}(Y)g(Y)], \quad \text{for all } g(Y) \in M.$$

Since boundedness of g implies $g(Y) \in M$, (10.7) holds.

When $X \in L^2$, we have shown that $\widehat{g}(Y)$ is the projection of X onto M. For $X \in L^1$, we proceed as follows. First consider the case of nonnegative X with $\mathsf{E}[X] < \infty$. We can approximate X by the bounded function X_n of Example 10.6. Being bounded, $X_n \in L^2$, and the corresponding $\widehat{g}_n(Y)$ exists and satisfies

$$\mathsf{E}[X_n g(Y)] \;=\; \mathsf{E}[\widehat{g}(Y)g(Y)], \quad \text{for all } g(Y) \in M. \tag{10.9}$$

Since $X_n \le X_{n+1}$, $\widehat{g}_n(Y) \le \widehat{g}_{n+1}(Y)$ by Problem 29. Hence, $\widehat{g}(Y) := \lim_{n\to\infty} \widehat{g}_n(Y)$ exists. To verify that $\widehat{g}(Y)$ satisfies (10.7), write[2]

$$\begin{aligned}
\mathsf{E}[Xg(Y)] &= \mathsf{E}\big[\lim_{n\to\infty} X_n g(Y)\big] \\
&= \lim_{n\to\infty} \mathsf{E}[X_n g(Y)] \\
&= \lim_{n\to\infty} \mathsf{E}[\widehat{g}_n(Y)g(Y)], \quad \text{by (10.9)}, \\
&= \mathsf{E}\big[\lim_{n\to\infty} \widehat{g}_n(Y)g(Y)\big] \\
&= \mathsf{E}[\widehat{g}(Y)g(Y)].
\end{aligned}$$

For signed X with $\mathsf{E}[|X|] < \infty$, consider the nonnegative random variables

$$X^+ \;:=\; \begin{cases} X, & X \ge 0, \\ 0, & X < 0, \end{cases} \quad \text{and} \quad X^- \;:=\; \begin{cases} \Leftrightarrow X, & X < 0, \\ 0, & X \ge 0. \end{cases}$$

[‡]See, for example, the derivation following (7.8) in Section 7.3.

Since $X^+ + X^- = |X|$, it is clear that X^+ and X^- are L^1 random variables. Since they are nonnegative, their conditional expectations exist. Denote them by $\widehat{g}^+(Y)$ and $\widehat{g}^-(Y)$. Since $X^+ \Leftrightarrow X^- = X$, it is easy to verify that $\widehat{g}(Y) := \widehat{g}^+(Y) \Leftrightarrow \widehat{g}^-(Y)$ satisfies (10.7).

Notation

We have shown that to every $X \in L^1$, there corresponds a unique function $\widehat{g}(y)$ such that (10.7) holds. The standard notation for this function of y is $\mathsf{E}[X|Y=y]$, which, as noted above, is given by the usual formulas when X and Y are discrete or jointly continuous. It is conventional in probability theory to write $\mathsf{E}[X|Y]$ instead of $\widehat{g}(Y)$. We emphasize that $\mathsf{E}[X|Y=y]$ is a deterministic function of y, while $\mathsf{E}[X|Y]$ is a function of Y and is therefore a random variable. We also point out that with the conventional notation, (10.7) becomes

$$\mathsf{E}[Xg(Y)] \;=\; \mathsf{E}\big[\mathsf{E}[X|Y]g(Y)\big], \quad \text{for all bounded functions } g. \tag{10.10}$$

10.6. The Spectral Representation

Let X_n be a discrete-time, zero-mean, wide-sense stationary process with correlation function

$$R(n) \;:=\; \mathsf{E}[X_{n+m}X_m]$$

and corresponding power spectral density

$$S(f) \;=\; \sum_{n=-\infty}^{\infty} R(n)e^{j2\pi fn} \tag{10.11}$$

so that[§]

$$R(n) \;=\; \int_{-1/2}^{1/2} S(f)e^{j2\pi fn}\,df. \tag{10.12}$$

Consider the space of complex-valued frequency functions,

$$\mathcal{S} \;:=\; \left\{G : \int_{-1/2}^{1/2} |G(f)|^2 S(f)\,df < \infty\right\}$$

equipped with the inner product

$$\langle G, H\rangle \;:=\; \int_{-1/2}^{1/2} G(f)H(f)^* S(f)\,df$$

[§]For some correlation functions, the sum in (10.11) may not converge. However, by **Herglotz's Theorem**, we can always replace (10.12) by

$$R(n) \;=\; \int_{-1/2}^{1/2} e^{j2\pi fn}\,dS_0(f),$$

where $S_0(f)$ is the spectral (cumulative) distribution function, and the rest of the section goes through with the necessary changes. Note that when the power spectral density exists, $S(f) = S_0'(f)$.

and corresponding norm $\|G\| := \langle G, G\rangle^{1/2}$. Then $\mathcal{S}$ is a Hilbert space. Furthermore, every $G \in \mathcal{S}$ can be approximated by a trigonometric polynomial of the form

$$G_0(f) \;=\; \sum_{n=-N}^{N} g_n e^{j2\pi fn}.$$

The approximation is in norm; i.e., given any $\varepsilon > 0$, there is a trigonometric polynomial G_0 such that $\|G \Leftrightarrow G_0\| < \varepsilon$ [5, p. 139].

To each frequency function $G \in \mathcal{S}$, we now associate an L^2 random variable as follows. For trigonometric polynomials like G_0, put

$$T(G_0) \;:=\; \sum_{n=-N}^{N} g_n X_n.$$

Note that T is well defined (Problem 35). A critical property of these trigonometric polynomials is that (Problem 36)

$$\mathsf{E}[T(G_0)T(H_0)^*] \;=\; \int_{-1/2}^{1/2} G_0(f)\,H_0(f)^*\,S(f)\,df \;=\; \langle G_0, H_0\rangle,$$

where H_0 is defined similarly to G_0. In particular, T is **norm preserving** on the trigonometric polynomials; i.e.,

$$\|T(G_0)\|_2^2 \;=\; \|G_0\|^2.$$

To define T for arbitrary $G \in \mathcal{S}$, let G_n be a sequence of trigonometric polynomials converging to G in norm; i.e., $\|G_n \Leftrightarrow G\| \to 0$. Then G_n is Cauchy (cf. Problem 11). Furthermore, the linearity of and norm-preservation properties of T on the trigonometric polynomials tells us that

$$\|G_n \Leftrightarrow G_m\| \;=\; \|T(G_n \Leftrightarrow G_m)\| \;=\; \|T(G_n) \Leftrightarrow T(G_m)\|_2,$$

and we see that $T(G_n)$ is Cauchy in L^2. Since L^2 is complete, there is a limit random variable, denoted by $T(G)$ and such that $\|T(G_n) \Leftrightarrow T(G)\|_2 \to 0$. Note that $T(G)$ is well defined, norm preserving, linear, and continuous on $\mathcal{S}$ (Problem 37).

There is another way approximate elements of $\mathcal{S}$. Every $G \in \mathcal{S}$ can also be approximated by a piecewise constant function of the following form. For $\Leftrightarrow 1/2 \le f_0 < \cdots < f_n \le 1/2$, let

$$G_0(f) \;=\; \sum_{i=1}^{n} g_i I_{(f_{i-1}, f_i]}(f).$$

Given any $\varepsilon > 0$, there is a piecewise constant function G_0 such that $\|G \Leftrightarrow G_0\| < \varepsilon$. This is exactly what we did for the Wiener process, but with a different norm. For piecewise constant G_0,

$$T(G_0) \;=\; \sum_{i=1}^{n} g_i T(I_{(f_{i-1}, f_i]}).$$

Since

$$I_{(f_{i-1},f_i]} \;=\; I_{[-1/2,f_i]} \Leftrightarrow I_{[-1/2,f_{i-1}]},$$

if we put

$$Z_f \;:=\; T(I_{[-1/2,f]}), \quad \Leftrightarrow 1/2 \le f \le 1/2,$$

then Z_f is well defined by Problem 38, and

$$T(G_0) \;=\; \sum_{i=1}^{n} g_i(Z_{f_i} \Leftrightarrow Z_{f_{i-1}}).$$

The family of random variables $\{Z_f, \Leftrightarrow 1/2 \le f \le 1/2\}$ has many similarities to the Wiener process (see Problem 39). We write

$$\int_{-1/2}^{1/2} G_0(f)\, dZ_f \;:=\; \sum_{i=1}^{n} g_i(Z_{f_i} \Leftrightarrow Z_{f_{i-1}}).$$

For arbitrary $G \in \mathcal{S}$, we approximate G with a sequence of piecewise constant functions. Then G_n will be Cauchy. Since

$$\left\| \int_{-1/2}^{1/2} G_n(f)\, dZ_f \Leftrightarrow \int_{-1/2}^{1/2} G_m(f)\, dZ_f \right\|_2 \;=\; \|T(G_n) \Leftrightarrow T(G_m)\|_2 \;=\; \|G_n \Leftrightarrow G_m\|,$$

there is a limit random variable in L^2, denoted by

$$\int_{-1/2}^{1/2} G(f)\, dZ_f,$$

and

$$\left\| \int_{-1/2}^{1/2} G_n(f)\, dZ_f \Leftrightarrow \int_{-1/2}^{1/2} G(f)\, dZ_f \right\|_2 \;\to\; 0.$$

On the other hand, since G_n is piecewise constant,

$$\int_{-1/2}^{1/2} G_n(f)\, dZ_f \;=\; T(G_n),$$

and since $\|G_n \Leftrightarrow G\| \to 0$, and since T is continuous, $\|T(G_n) \Leftrightarrow T(G)\|_2 \to 0$. Thus,

$$T(G) \;=\; \int_{-1/2}^{1/2} G(f)\, dZ_f.$$

In particular, taking $G(f) = e^{j2\pi fn}$ shows that

$$\int_{-1/2}^{1/2} e^{j2\pi fn}\, dZ_f \;=\; T(G) \;=\; X_n,$$

where the last step follows from the original definition of T. The formula

$$X_n \;=\; \int_{-1/2}^{1/2} e^{j2\pi fn}\, dZ_f$$

is called the **spectral representation** of X_n.

10.7. Notes

Notes §10.1: Convergence in Mean of Order p

***Note* 1.** The assumption

$$\sum_{k=1}^{\infty} |h_k| \;<\; \infty$$

should be understood more precisely as saying that the partial sum

$$A_n \;:=\; \sum_{k=1}^{n} |h_k|$$

is bounded above by some finite constant. Since A_n is also nondecreasing, the **monotonic sequence property** of real numbers [38, p. 55, Theorem 3.14] says that A_n converges to a real number, say A. Now, by the same argument used to solve Problem 11, since A_n converges, it is Cauchy. Hence, given any $\varepsilon > 0$, for large enough n and m, $|A_n \Leftrightarrow A_m| < \varepsilon$. If $n > m$, we have

$$A_n \Leftrightarrow A_m \;=\; \sum_{k=m+1}^{n} |h_k| \;<\; \varepsilon.$$

Notes §10.5: Conditional Expectation

***Note* 2.** The interchange of limit and expectation is justified by Lebesgue's monotone convergence theorem [4, p. 208].

10.8. Problems

Problems §10.1: Convergence in Mean of Order p

1. Let $U \sim \text{uniform}[0,1]$, and put

$$X_n \;:=\; \sqrt{n} I_{[0,1/n]}(U), \quad n = 1, 2, \ldots.$$

   For what values of $p \geq 1$ does X_n converge in mean of order p to zero?

2. Let N_t be a Poisson process of rate λ. Show that N_t/t converges in mean square to λ as $t \to \infty$.

3. Show that $\lim_{k\to\infty} R(k) = 0$ implies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} R(k) \;=\; 0,$$

   and thus M_n converges in mean square to m by Example 10.4.

4. Show that if M_n converges in mean square m, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} R(k) \;=\; 0,$$

where $R(k)$ is defined in Example 10.4. *Hint:* Observe that

$$\begin{aligned} \sum_{k=0}^{n-1} R(k) &= \mathsf{E}\left[(X_1 \Leftrightarrow m) \sum_{k=1}^{n} (X_k \Leftrightarrow m)\right] \\ &= \mathsf{E}[(X_1 \Leftrightarrow m) \cdot n(M_n \Leftrightarrow m)]. \end{aligned}$$

5. Let Z be a nonnegative random variable with $\mathsf{E}[Z] < \infty$. Given any $\varepsilon > 0$, show that there is a $\delta > 0$ such that for any event A,

$$\wp(A) \;<\; \delta \quad \text{implies} \quad \mathsf{E}[ZI_A] \;<\; \varepsilon.$$

Hint: Recalling Example 10.6, put $Z_n := \min(Z, n)$ and write

$$\mathsf{E}[ZI_A] \;=\; \mathsf{E}[(Z \Leftrightarrow Z_n) I_A] + \mathsf{E}[Z_n I_A].$$

6. Derive **Hölder's inequality**,

$$\mathsf{E}[|XY|] \;\le\; \mathsf{E}[|X|^p]^{1/p} \mathsf{E}[|Y|^q]^{1/q},$$

if $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. *Hint:* Let α and β denote the factors on the right-hand side; i.e., $\alpha := \mathsf{E}[|X|^p]^{1/p}$ and $\beta := \mathsf{E}[|Y|^q]^{1/q}$. Then it suffices to show that

$$\frac{\mathsf{E}[|XY|]}{\alpha\beta} \;\le\; 1.$$

To this end, observe that by the convexity of the exponential function, for any real numbers u and v, we can always write

$$\exp\left[\tfrac{1}{p}u + \tfrac{1}{q}v\right] \;\le\; \tfrac{1}{p}e^u + \tfrac{1}{q}e^v.$$

Now take $u = \ln(|X|/\alpha)^p$ and $v = \ln(|Y|/\beta)^q$.

7. Derive **Lyapunov's inequality**,

$$\mathsf{E}[|Z|^\alpha]^{1/\alpha} \;\le\; \mathsf{E}[|Z|^\beta]^{1/\beta}, \quad 1 \le \alpha < \beta < \infty.$$

Hint: Apply Hölder's inequality to $X = |Z|^\alpha$ and $Y = 1$ with $p = \beta/\alpha$.

8. Show that if X_n converges in mean of order $\beta > 1$ to X, then X_n converges in mean of order α to X for all $1 \le \alpha < \beta$.

9. Derive **Minkowski's inequality**,

$$\mathsf{E}[|X+Y|^p]^{1/p} \;\leq\; \mathsf{E}[|X|^p]^{1/p} + \mathsf{E}[|Y|^p]^{1/p},$$

where $1 \leq p < \infty$. *Hint:* Observe that

$$\begin{aligned}\mathsf{E}[|X+Y|^p] &= \mathsf{E}[|X+Y|\,|X+Y|^{p-1}] \\ &\leq \mathsf{E}[|X|\,|X+Y|^{p-1}] + \mathsf{E}[|Y|\,|X+Y|^{p-1}],\end{aligned}$$

and apply Hölder's inequality.

Problems §10.2: Normed Vector Spaces of Random Variables

10. Show that if X_n converges in mean of order p to X, and if Y_n converges in mean of order p to Y, then $X_n + Y_n$ converges in mean of order p to $X+Y$.

11. Let X_n converge in mean of order p to $X \in L^p$. Show that X_n is Cauchy.

12. Let X_n be a Cauchy sequence in L^p. Show that X_n is bounded in the sense that there is a finite constant K such that $\|X_n\|_p \leq K$ for all n.

    Remark. Since by the previous problem, a convergent sequence is Cauchy, it follows that a convergent sequence is bounded.

13. Let X_n converge in mean of order p to $X \in L^p$, and let Y_n converge in mean of order p to $Y \in L^p$. Show that X_nY_n converges in mean of order p to XY.

14. Show that limits in mean of order p are unique; i.e., if X_n converges in mean of order p to both X and Y, show that $\mathsf{E}[|X \ominus Y|^p] = 0$.

15. Show that

$$\big|\|X\|_p \ominus \|Y\|_p\big| \;\leq\; \|X \ominus Y\|_p \;\leq\; \|X\|_p + \|Y\|_p\,.$$

16. If X_n converges in mean of order p to X, show that

$$\lim_{n\to\infty} \mathsf{E}[|X_n|^p] \;=\; \mathsf{E}[|X|^p].$$

17. Show that

$$\sum_{k=1}^{\infty} h_k\, X_k$$

is well defined as an element of L^1 assuming that

$$\sum_{k=1}^{\infty} |h_k| \;<\; \infty \quad \text{and} \quad \mathsf{E}[|X_k|] \;\leq\; B, \quad \text{for all } k,$$

where B is a finite constant.

Remark. By the Cauchy–Schwarz inequality, a sufficient condition for $\mathsf{E}[|X_k|] \le B$ is that $\mathsf{E}[X_k^2] \le B^2$.

Problems §10.3: The Wiener Integral (Again)

18. Recall that the Wiener integral of g was defined to be the mean-square limit of

$$Y_n \;:=\; \int_0^\infty g_n(\tau)\, dW_\tau,$$

where each g_n is piecewise constant, and $\|g_n \Leftrightarrow g\| \to 0$. Show that

$$\mathsf{E}\left[\int_0^\infty g(\tau)\, dW_\tau\right] \;=\; 0,$$

and

$$\mathsf{E}\left[\left(\int_0^\infty g(\tau)\, dW_\tau\right)^2\right] \;=\; \sigma^2 \int_0^\infty g(\tau)^2\, d\tau.$$

Hint: Let Y denote the mean-square limit of Y_n. Show that $\mathsf{E}[Y_n] \to \mathsf{E}[Y]$ and that $\mathsf{E}[Y_n^2] \to \mathsf{E}[Y^2]$.

19. Use the following approach to show that the limit definition of the Wiener integral is unique. Let Y_n and Y be as in the text. If $\widetilde{g}_n$ is another sequence of piecewise constant functions with $\|\widetilde{g}_n \Leftrightarrow g\|^2 \to 0$, put

$$\widetilde{Y}_n \;:=\; \int_0^\infty \widetilde{g}_n(\tau)\, dW_\tau.$$

By the argument that Y_n has a limit Y, $\widetilde{Y}_n$ has a limit, say $\widetilde{Y}$. Show that $\|Y \Leftrightarrow \widetilde{Y}\|_2 = 0$.

20. Show that the Wiener integral is linear even for functions that are not piecewise constant.

Problems §10.4: Projections, Orthogonality Principle, Projection Theorem

21. Let $C = \{Y \in L^p : \|Y\|_p \le r\}$. For $\|X\|_p > r$, find a projection of X onto C.

22. Show that if $\widehat{X} \in M$ satisfies (10.6), it is unique.

23. Let $N \subset M$ be subspaces of L^2. Assume that the projection of $X \in L^2$ onto M exists and denote it by $\widehat{X}_M$. Similarly, assume that the projection of X onto N exists and denote it by $\widehat{X}_N$. Show that $\widehat{X}_N$ is the projection of $\widehat{X}_M$ onto N.

24. The preceding problem shows that when $N \subset M$ are subspaces, the projection of X onto N can be computed in two stages: First project X onto M and then project the result onto N. Show that this is not true in general if N and M are not both subspaces. *Hint:* Draw a disk of radius one. Then draw a straight line through the origin. Identify M with the disk and identify N with the line segment obtained by intersecting the line with the disk. If the point X is outside the disk and not on the line, then projecting first onto the ball and then onto the segment does not give the projection.

    Remark. Interestingly, projecting first onto the line (not the line segment) and then onto the disk does give the correct answer.

25. Derive the **parallelogram law**,

$$\|X+Y\|_2^2 + \|X \Leftrightarrow Y\|_2^2 \;=\; 2(\|X\|_2^2 + \|Y\|_2^2).$$

26. Show that the result of Example 10.7 holds if the assumption $\sum_{k=1}^{\infty} |h_k| < \infty$ is replaced by the two assumptions $\sum_{k=1}^{\infty} |h_k|^2 < \infty$ and $\mathsf{E}[X_k X_l] = 0$ for $k \neq l$.

27. If $\|X_n \Leftrightarrow X\|_2 \to 0$ and $\|Y_n \Leftrightarrow Y\|_2 \to 0$, show that

$$\lim_{n\to\infty} \langle X_n, Y_n \rangle \;=\; \langle X, Y \rangle.$$

28. If Y has density f_Y, show that M in (10.8) is closed. *Hints:* (*i*) Observe that

$$\mathsf{E}[g(Y)^2] \;=\; \int_{-\infty}^{\infty} g(y)^2 f_Y(y)\, dy.$$

    (*ii*) Use the fact that the set of functions

$$G \;:=\; \left\{ g : \int_{-\infty}^{\infty} g(y)^2 f_Y(y)\, dy < \infty \right\}$$

    is complete if G is equipped with the norm

$$\|g\|_Y^2 \;:=\; \int_{-\infty}^{\infty} g(y)^2 f_Y(y)\, dy.$$

    (*iii*) Follow the method of Example 10.9.

Problems §10.5: Conditional Expectation

29. Fix $X \in L^1$, and suppose $X \geq 0$. Show that $\mathsf{E}[X|Y] \geq 0$ in the sense that $\wp(\mathsf{E}[X|Y] < 0) = 0$. *Hints:* To begin, note that

$$\{\mathsf{E}[X|Y] < 0\} \;=\; \bigcup_{n=1}^{\infty} \{\mathsf{E}[X|Y] < \Leftrightarrow 1/n\}.$$

By limit property (1.4),

$$\wp(\mathsf{E}[X|Y] < 0) \;=\; \lim_{n\to\infty} \wp(\mathsf{E}[X|Y] < \Leftrightarrow 1/n).$$

To obtain a proof by contradiction, suppose that $\wp(\mathsf{E}[X|Y] < 0) > 0$. Then for sufficiently large n, $\wp(\mathsf{E}[X|Y] < \Leftrightarrow 1/n) > 0$ too. Take $g(Y) = I_B(\mathsf{E}[X|Y])$ where $B = (\Leftrightarrow\infty, \Leftrightarrow 1/n)$ and consider the defining relationship

$$\mathsf{E}[Xg(Y)] \;=\; \mathsf{E}\big[\mathsf{E}[X|Y]g(Y)\big].$$

30. For $X \in L^1$, let X^+ and X^- be as in the text, and denote their corresponding conditional expectations by $\mathsf{E}[X^+|Y]$ and $\mathsf{E}[X^-|Y]$, respectively. Show that

$$\mathsf{E}[Xg(Y)] \;=\; \mathsf{E}\big[\big(\mathsf{E}[X^+|Y] \Leftrightarrow \mathsf{E}[X^-|Y]\big)g(Y)\big].$$

31. For $X \in L^1$, derive the **smoothing property** of conditional expectation,

$$\mathsf{E}[X|Y_1] \;=\; \mathsf{E}\big[\mathsf{E}[X|Y_2, Y_1]\big|Y_1\big].$$

Remark. If $X \in L^2$, this is an instance of Problem 23.

32. If $X \in L^1$ and h is a bounded function, show that

$$\mathsf{E}[h(Y)X|Y] \;=\; h(Y)\mathsf{E}[X|Y].$$

33. If $X \in L^2$ and $h(Y) \in L^2$, use the orthogonality principle to show that

$$\mathsf{E}[h(Y)X|Y] \;=\; h(Y)\mathsf{E}[X|Y].$$

34. If $X \in L^1$ and if $h(Y)X \in L^1$, show that

$$\mathsf{E}[h(Y)X|Y] \;=\; h(Y)\mathsf{E}[X|Y].$$

Hint: Use a limit argument as in the text to approximate $h(Y)$ by a bounded function of Y. Then apply either of the preceding two problems.

Problems §10.6: The Spectral Representation

35. Show that the mapping T defined by

$$G_0(f) \;=\; \sum_{n=-N}^{N} g_n e^{j2\pi fn} \;\mapsto\; \sum_{n=-N}^{N} g_n X_n$$

is well defined; i.e., if $G_0(f)$ has another representation, say

$$G_0(f) \;=\; \sum_{n=-N}^{N} \widetilde{g}_n e^{j2\pi fn},$$

show that

$$\sum_{n=-N}^{N} \widetilde{g}_n X_n \;=\; \sum_{n=-N}^{N} g_n X_n.$$

Hint: With $d_n := g_n \Leftrightarrow \widetilde{g}_n$, it suffices to show that if $\sum_{n=-N}^{N} d_n e^{j2\pi fn} = 0$, then $Y := \sum_{n=-N}^{N} d_n X_n = 0$, and for this it is enough to show that $\mathsf{E}[|Y|^2] = 0$. Formula (10.12) may be helpful.

36. Show that if $H_0(f)$ is defined similarly to $G_0(f)$ in the preceding problem, then

$$\mathsf{E}[T(G_0)T(H_0)^*] \;=\; \int_{-1/2}^{1/2} G_0(f) H_0(f)^* S(f)\, df.$$

37. Show that if G_n and $\widetilde{G}_n$ are trigonometric polynomials with $\|G_n \Leftrightarrow G\| \to 0$ and $\|\widetilde{G}_n \Leftrightarrow G\| \to 0$, then

$$\lim_{n\to\infty} T(G_n) \;=\; \lim_{n\to\infty} T(\widetilde{G}_n).$$

Also show that T is norm preserving, linear, and continuous on $\mathcal{S}$. *Hint:* Put $T(G) := \lim_{n\to\infty} T(G_n)$ and $Y := \lim_{n\to\infty} T(\widetilde{G}_n)$. Show that $\|T(G) \Leftrightarrow Y\|_2 = 0$. Use the fact that T is norm preserving on trigonometric polynomials.

38. Show that $Z_f := T(I_{[-1/2,f]})$ is well defined. *Hint:* It suffices to show that $I_{[-1/2,f]} \in \mathcal{S}$.

39. Since

$$\int_{-1/2}^{1/2} G(f)\, dZ_f \;=\; T(G),$$

use results about $T(G)$ to prove the following.

(a) Show that

$$\mathsf{E}\left[\int_{-1/2}^{1/2} G(f)\, dZ_f\right] \;=\; 0.$$

(b) Show that

$$\mathsf{E}\left[\left(\int_{-1/2}^{1/2} G(f)\, dZ_f\right)\left(\int_{-1/2}^{1/2} H(\nu)\, dZ_\nu\right)^*\right] \;=\; \int_{-1/2}^{1/2} G(f) H(f)^* S(f)\, df.$$

(c) Show that Z_f has orthogonal (uncorrelated) increments.

CHAPTER 11

Other Modes of Convergence

This chapter introduces the three types of convergence,

- convergence in probability,
- convergence in distribution,
- almost sure convergence,

which are related to each other (and to convergence in mean of order p) as shown in Figure 11.1.

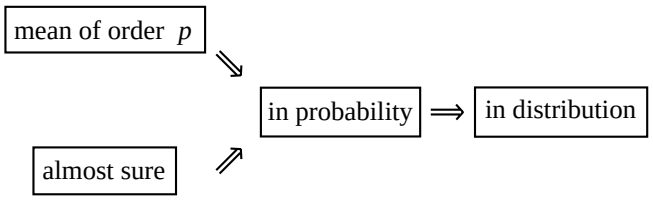

Figure 11.1. Implications of various types of convergence.

Section 11.1 introduces the notion of convergence in probability. Convergence in probability will be important in Chapter 12 on parameter estimation and confidence intervals, where it justifies various statistical procedures that are used to estimate unknown parameters.

Section 11.2 introduces the notion of convergence in distribution. Convergence in distribution is often used to approximate probabilities that are hard to calculate exactly. Suppose that X_n is a random variable whose cumulative distribution function $F_{X_n}(x)$ is very hard to compute. But suppose that for large n, $F_{X_n}(x) \approx F_X(x)$, where $F_X(x)$ is a cdf that is very easy to compute. When $F_{X_n}(x) \to F_X(x)$, we say that X_n converges in distribution to X. In this case, we can approximate $F_{X_n}(x)$ by $F_X(x)$ if n is large enough. When the central limit theorem applies, F_X is the normal cdf with mean zero and variance one.

Section 11.3 introduces the notion of almost-sure convergence. This kind of convergence is more technically demanding to analyze, but it allows us to discuss important results such as the strong law of large numbers, for which we give the usual fourth-moment proof. Almost-sure convergence also allows us the derive the Skorohod representation, which is a powerful tool for studying convergence in distribution.

11.1. Convergence in Probability

We say that X_n **converges in probability** to X if

$$\lim_{n\to\infty} \wp(|X_n \Leftrightarrow X| \geq \varepsilon) \;=\; 0 \quad \text{for all } \varepsilon > 0.$$

The first thing to note about convergence in probability is that it is implied by convergence in mean of order p. Using Markov's inequality, we have

$$\wp(|X_n \Leftrightarrow X| \geq \varepsilon) \;=\; \wp(|X_n \Leftrightarrow X|^p \geq \varepsilon^p) \;\leq\; \frac{\mathsf{E}[|X_n \Leftrightarrow X|^p]}{\varepsilon^p}.$$

Hence, if X_n converges in mean of order p to X, then X_n converges in probability to X as well.

***Example* 11.1** (Weak Law of Large Numbers). Since the sample mean M_n defined in Example 10.3 converges in mean square to m, it follows that M_n converges in probability to m. This is known as the weak law of large numbers.

The second thing to note about convergence in probability is that it is possible to have X_n converging in probability to X, while X_n does not converge to X in mean of order p for any $p \geq 1$. For example, let $U \sim$ uniform$[0,1]$, and consider the sequence

$$X_n \;:=\; nI_{[0,1/n]}(U), \quad n = 1, 2, \ldots.$$

Observe that for $\varepsilon > 0$,

$$\{|X_n| \geq \varepsilon\} \;=\; \begin{cases} \{U \leq 1/n\}, & n \geq \varepsilon, \\ \emptyset, & n < \varepsilon. \end{cases}$$

It follows that for all n,

$$\wp(|X_n| \geq \varepsilon) \;\leq\; \wp(U \leq 1/n) \;=\; 1/n \;\to\; 0.$$

Thus, X_n converges in probability to zero. However,

$$\mathsf{E}[|X_n|^p] \;=\; n^p \wp(U \leq 1/n) \;=\; n^{p-1},$$

which does not go to zero as $n \to \infty$.

The third thing to note about convergence in probability is that if $g(x,y)$ is continuous, and if X_n converges in probability to X, and Y_n converges in probability to Y, then $g(X_n, Y_n)$ converges in probability to $g(X,Y)$. This result is derived in Problem 6. In particular, note that since the functions $g(x,y) = x + y$ and $g(x,y) = xy$ are continuous, $X_n + Y_n$ and $X_n Y_n$ converge in probability to $X + Y$ and XY, respectively, whenever X_n and Y_n converge in probability to X and Y, respectively.

***Example* 11.2.** Since Problem 6 is somewhat involved, it is helpful to first see the derivation for the special case in which X_n and Y_n both converge in probability to *constant* random variables, say u and v, respectively.

Solution. Let $\varepsilon > 0$ be given. We show that

$$\wp(|g(X_n, Y_n) \Leftrightarrow g(u, v)| \geq \varepsilon) \;\to\; 0.$$

By the continuity of g at the point (u, v), there is a $\delta > 0$ such that whenever (u', v') is within δ of (u, v); i.e., whenever

$$\sqrt{|u' \Leftrightarrow u|^2 + |v' \Leftrightarrow v|^2} \;<\; 2\delta,$$

then

$$|g(u', v') \Leftrightarrow g(u, v)| \;<\; \varepsilon.$$

Since

$$\sqrt{|u' \Leftrightarrow u|^2 + |v' \Leftrightarrow v|^2} \;\leq\; |u' \Leftrightarrow u| + |v' \Leftrightarrow v|,$$

we see that

$$|X_n \Leftrightarrow u| < \delta \text{ and } |Y_n \Leftrightarrow v| < \delta \;\Rightarrow\; |g(X_n, Y_n) \Leftrightarrow g(u, v)| < \varepsilon.$$

Conversely,

$$|g(X_n, Y_n) \Leftrightarrow g(u, v)| \geq \varepsilon \;\Rightarrow\; |X_n \Leftrightarrow u| \geq \delta \text{ or } |Y_n \Leftrightarrow v| \geq \delta.$$

It follows that

$$\wp(|g(X_n, Y_n) \Leftrightarrow g(u, v)| \geq \varepsilon) \leq \wp(|X_n \Leftrightarrow u| \geq \delta) + \wp(|Y_n \Leftrightarrow v| \geq \delta).$$

These last two terms go to zero by hypothesis.

***Example* 11.3.** Let X_n converge in probability to x, and let c_n be a converging sequence of real numbers with limit c. Show that $c_n X_n$ converges in probability to cx.

Solution. Define the sequence of constant random variables $Y_n := c_n$. It is a simple exercise to show that Y_n converges in probability to c. Thus, $X_n Y_n = c_n X_n$ converges in probability to cx.

11.2. Convergence in Distribution

We say that X_n **converges in distribution** to X, or **converges weakly** to X if

$$\lim_{n\to\infty} F_{X_n}(x) \;=\; F_X(x), \quad \text{for all } x \in C(F_X),$$

where $C(F_X)$ is the set of points x at which F_X is continuous. If F_X has a jump at a point x_0, then we do not say anything about the existence of

$$\lim_{n\to\infty} F_{X_n}(x_0).$$

The first thing to note about convergence in distribution is that it is implied by convergence in probability. To derive this result, fix any x at which F_X is continuous. For any $\varepsilon > 0$, write

$$\begin{aligned} F_{X_n}(x) &= \wp(X_n \le x, X \le x+\varepsilon) + \wp(X_n \le x, X > x+\varepsilon) \\ &\le F_X(x+\varepsilon) + \wp(X_n \Leftrightarrow X < \Leftrightarrow\varepsilon) \\ &\le F_X(x+\varepsilon) + \wp(|X_n \Leftrightarrow X| \ge \varepsilon). \end{aligned}$$

It follows that

$$\overline{\lim_{n\to\infty}} F_{X_n}(x) \le F_X(x+\varepsilon).$$

Similarly,

$$\begin{aligned} F_X(x \Leftrightarrow\varepsilon) &= \wp(X \le x \Leftrightarrow\varepsilon, X_n \le x) + \wp(X \le x \Leftrightarrow\varepsilon, X_n > x) \\ &\le F_{X_n}(x) + \wp(X \Leftrightarrow X_n < \Leftrightarrow\varepsilon) \\ &\le F_{X_n}(x) + \wp(|X_n \Leftrightarrow X| \ge \varepsilon), \end{aligned}$$

and we obtain

$$F_X(x \Leftrightarrow\varepsilon) \le \underline{\lim_{n\to\infty}} F_{X_n}(x).$$

Since the liminf is always less than or equal to the limsup, we have

$$F_X(x \Leftrightarrow\varepsilon) \le \underline{\lim_{n\to\infty}} F_{X_n}(x) \le \overline{\lim_{n\to\infty}} F_{X_n}(x) \le F_X(x+\varepsilon).$$

Since F_X is continuous at x, letting ε go to zero shows that the liminf and the limsup are equal to each other and to $F_X(x)$. Hence $\lim_{n\to\infty} F_{X_n}(x)$ exists and equals $F_X(x)$.

The need to restrict attention to continuity points can also be seen in the following example.

***Example* 11.4.** Let $X_n \sim \exp(n)$; i.e.,

$$F_{X_n}(x) = \begin{cases} 1 \Leftrightarrow e^{-nx}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

For $x > 0$, $F_{X_n}(x) \to 1$ as $n \to \infty$. However, since $F_{X_n}(0) = 0$, we see that $F_{X_n}(x) \to I_{(0,\infty)}(x)$, which, being left continuous at zero, is not the cumulative distribution function of any random variable. Fortunately, the constant random variable $X \equiv 0$ has $I_{[0,\infty)}(x)$ for its cdf. Since the only point of discontinuity is zero, $F_{X_n}(x)$ does indeed converge to $F_X(x)$ at all continuity points of F_X.

The second thing to note about convergence in distribution is that if X_n converges in distribution to X, and if X is a *constant* random variable, say $X \equiv c$, then X_n converges in probability to c. To derive this result, first observe that

$$\begin{aligned}\{|X_n \ominus c| < \varepsilon\} &= \{\ominus\varepsilon < X_n \ominus c < \varepsilon\} \\ &= \{c \ominus \varepsilon < X_n\} \cap \{X_n < c + \varepsilon\}.\end{aligned}$$

It is then easy to see that

$$\begin{aligned}\wp(|X_n \ominus c| \geq \varepsilon) &\leq \wp(X_n \leq c \ominus \varepsilon) + \wp(X_n \geq c + \varepsilon) \\ &\leq F_{X_n}(c \ominus \varepsilon) + \wp(X_n > c + \varepsilon/2) \\ &= F_{X_n}(c \ominus \varepsilon) + 1 \ominus F_{X_n}(c + \varepsilon/2).\end{aligned}$$

Since $X \equiv c$, $F_X(x) = I_{[c,\infty)}(x)$. Therefore, $F_{X_n}(c \ominus \varepsilon) \to 0$ and $F_{X_n}(c+\varepsilon/2) \to 1$.

The third thing to note about convergence in distribution is that it is equivalent to the condition

$$\lim_{n\to\infty} \mathsf{E}[g(X_n)] = \mathsf{E}[g(X)] \quad \text{for every bounded continuous function } g. \quad (11.1)$$

A proof that convergence in distribution implies (11.1) can be found in [17, p. 316]. A proof that (11.1) implies convergence in distribution is sketched in Problem 19.

***Example* 11.5.** Show that if X_n converges in distribution to X, then the characteristic function of X_n converges to the characteristic function of X.

Solution. Fix any ν, and take $g(x) = e^{j\nu x}$. Then

$$\varphi_{X_n}(\nu) = \mathsf{E}[e^{j\nu X_n}] = \mathsf{E}[g(X_n)] \to \mathsf{E}[g(X)] = \mathsf{E}[e^{j\nu X}] = \varphi_X(\nu).$$

Remark. It is also true that if $\varphi_{X_n}(\nu)$ converges to $\varphi_X(\nu)$ for all ν, then X_n converges in distribution to X [4, p. 349, Theorem 26.3].

***Example* 11.6.** Let $X_n \sim N(m_n, \sigma_n^2)$, where $m_n \to m$ and $\sigma_n^2 \to \sigma^2$. If $X \sim N(m, \sigma^2)$, show that X_n converges in distribution to X.

Solution. It suffices to show that the characteristic function of X_n converges to the characteristic function of X. Write

$$\varphi_{X_n}(\nu) = e^{jm_n\nu - \sigma_n^2\nu^2/2} \to e^{jm\nu - \sigma^2\nu^2/2} = \varphi_X(\nu).$$

A very important instance of convergence in distribution is the **central limit theorem**. This result says that if $X_1, X_2, \ldots$ are independent, identically distributed random variables with finite mean m and finite variance σ^2, and if

$$M_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad Y_n \;:=\; \frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}},$$

then

$$\lim_{n\to\infty} F_{Y_n}(y) \;=\; \Phi(y) \;:=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-t^2/2}\,dt.$$

In other words, for all y, $F_{Y_n}(y)$ converges to the standard normal cdf $\Phi(y)$ (which is continuous at all points y). To better see the difference between the weak law of large numbers and the central limit theorem, we specialize to the case $m = 0$ and $\sigma^2 = 1$. In this case,

$$Y_n \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i.$$

In other words, the sample mean M_n divides the sum $X_1 + \cdots + X_n$ by n while Y_n divides the sum only by $\sqrt{n}$. The difference is that M_n converges in probability (and in distribution) to the *constant* zero, while Y_n converges in distribution to an $N(0,1)$ *random variable*. Notice also that the weak law requires only uncorrelated random variables, while the central limit theorem requires i.i.d. random variables. The central limit theorem was derived in Section 4.7, where examples and problems can also be found. The central limit theorem is also used to determine confidence intervals in Chapter 12.

***Example* 11.7.** Let N_t be a Poisson process of rate λ. Show that

$$Y_n \;:=\; \frac{\frac{N_n}{n} \Leftrightarrow \lambda}{\sqrt{\lambda/n}}$$

converges in distribution to an $N(0,1)$ random variable.

Solution. Since $N_0 \equiv 0$,

$$\frac{N_n}{n} \;=\; \frac{1}{n}\sum_{k=1}^{n}(N_k \Leftrightarrow N_{k-1}).$$

By the independent increments property of the Poisson process, the terms of the sum are i.i.d. Poisson(λ) random variables, with mean λ and variance λ. Hence, Y_n has the structure to apply the central limit theorem, and so $F_{Y_n}(y) \to \Phi(y)$ for all y.

The next example is a version of **Slutsky's Theorem**. We need it in our analysis of confidence intervals in Chapter 12.

***Example* 11.8.** Let Y_n be a sequence of random variables with corresponding cdfs F_n. Suppose that F_n converges to a continuous cdf F. Suppose also that U_n converges in probability to 1. Show that

$$\lim_{n\to\infty} \wp(Y_n \le yU_n) \;=\; F(y).$$

Solution. The result is very intuitive. For large n, $U_n \approx 1$ and $F_n(y) \approx F(y)$ suggest that

$$\wp(Y_n \le yU_n) \;\approx\; \wp(Y_n \le y) \;=\; F_n(y) \;\approx\; F(y).$$

The precise details are more involved. Fix any $y > 0$ and $0 < \delta < 1$. Then $\wp(Y_n \le yU_n)$ is equal to

$$\wp(Y_n \le yU_n\,, |U_n \Leftrightarrow 1| < \delta) + \wp(Y_n \le yU_n\,, |U_n \Leftrightarrow 1| \ge \delta).$$

The second term is upper bounded by $\wp(|U_n \Leftrightarrow 1| \ge \delta)$, which goes to zero. Rewrite the first term as

$$\wp(Y_n \le yU_n\,, 1 \Leftrightarrow \delta < U_n < 1 + \delta),$$

which is equivalent to

$$\wp(Y_n \le yU_n\,, y(1 \Leftrightarrow \delta) < yU_n < y(1 + \delta)). \tag{11.2}$$

Now this is upper bounded by

$$\wp(Y_n \le y(1 + \delta)) \;=\; F_n(y(1 + \delta)).$$

Thus,

$$\underline{\lim}_{n\to\infty} \wp(Y_n \le yU_n) \;\le\; F(y(1 + \delta)).$$

Next, (11.2) is lower bounded by

$$\wp(Y \le y(1 \Leftrightarrow \delta), |U_n \Leftrightarrow 1| < \delta),$$

which is equal to

$$\wp(Y_n \le y(1 \Leftrightarrow \delta)) \Leftrightarrow \wp(Y_n \le y(1 \Leftrightarrow \delta), |U_n \Leftrightarrow 1| \ge \delta).$$

Now the second term satisfies

$$\wp(Y_n \le y(1 \Leftrightarrow \delta), |U_n \Leftrightarrow 1| \ge \delta) \;\le\; \wp(|U_n \Leftrightarrow 1| \ge \delta) \;\to\; 0.$$

In light of these observations,

$$\overline{\lim}_{n\to\infty} \wp(Y_n \le yU_n) \;\ge\; F(y(1 \Leftrightarrow \delta)).$$

Since δ was arbitrary, and since F is continuous,

$$\lim_{n\to\infty} \wp(Y_n \le yU_n) \;=\; F(y).$$

The case $y < 0$ is similar.

11.3. Almost Sure Convergence

Let X_n be any sequence of random variables, and let X be any other random variable. Put

$$G \;:=\; \big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\big\}.$$

In other words, G is the set of sample points $\omega \in \Omega$ for which the sequence of real numbers $X_n(\omega)$ converges to the real number $X(\omega)$. We think of G as the set of "good" ω's for which $X_n(\omega) \to X(\omega)$. Similarly, we think of the complement of G, G^c, as the set of "bad" ω's for which $X_n(\omega) \not\to X(\omega)$.

We say that X_n **converges almost surely** to X if[1]

$$\wp\big(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}\big) \;=\; 0. \tag{11.3}$$

In other words, X_n converges almost surely to X if the "bad" set G^c has probability zero. We write $X_n \to X$ a.s. to indicate that X_n converges almost surely to X.

If it should happen that the bad set $G^c = \emptyset$, then $X_n(\omega) \to X(\omega)$ for every $\omega \in \Omega$. This is called **sure convergence**, and is a special case of almost sure convergence.

Because almost sure convergence is so closely linked to the convergence of ordinary sequences of real numbers, many results are easy to derive.

***Example* 11.9.** Show that if $X_n \to X$ a.s. and $Y_n \to Y$ a.s., then $X_n + Y_n \to X + Y$ a.s.

Solution. Let $G_X := \{X_n \to X\}$ and $G_Y := \{Y_n \to Y\}$. In other words, G_X and G_Y are the "good" sets for the sequences X_n and Y_n respectively. Now consider any $\omega \in G_X \cap G_Y$. For such ω, the sequence of real numbers $X_n(\omega)$ converges to the real number $X(\omega)$, and the sequence of real numbers $Y_n(\omega)$ converges to the real number $Y(\omega)$. Hence, from convergence theory for sequences of real numbers,

$$X_n(\omega) + Y_n(\omega) \;\to\; X(\omega) + Y(\omega). \tag{11.4}$$

At this point, we have shown that

$$G_X \cap G_Y \;\subset\; G, \tag{11.5}$$

where G denotes the set of all ω for which (11.4) holds. To prove that $X_n + Y_n \to X + Y$ a.s., we must show that $\wp(G^c) = 0$. On account of (11.5), $G^c \subset G_X^c \cup G_Y^c$. Hence,

$$\wp(G^c) \;\le\; \wp(G_X^c) + \wp(G_Y^c),$$

and the two terms on the right are zero because X_n and Y_n both converge almost surely.

In order to discuss almost sure convergence in more detail, it is helpful to characterize when (11.3) holds. Recall that a sequence of real numbers x_n converges to the real number x if given any $\varepsilon > 0$, there is a positive integer N such that for all $n \geq N$, $|x_n \Leftrightarrow x| < \varepsilon$. Hence,

$$\{X_n \to X\} \;=\; \bigcap_{\varepsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|X_n \Leftrightarrow X| < \varepsilon\}.$$

Equivalently,

$$\{X_n \not\to X\} \;=\; \bigcup_{\varepsilon>0} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{|X_n \Leftrightarrow X| \geq \varepsilon\}.$$

It is convenient to put

$$A_n(\varepsilon) \;:=\; \{|X_n \Leftrightarrow X| \geq \varepsilon\}, \quad B_N(\varepsilon) \;:=\; \bigcup_{n=N}^{\infty} A_n(\varepsilon),$$

and

$$A(\varepsilon) \;:=\; \bigcap_{N=1}^{\infty} B_N(\varepsilon). \tag{11.6}$$

Then

$$\wp(\{X_n \not\to X\}) \;=\; \wp\Big(\bigcup_{\varepsilon>0} A(\varepsilon)\Big).$$

If $X_n \to X$ a.s., then

$$0 \;=\; \wp(\{X_n \not\to X\}) \;=\; \wp\Big(\bigcup_{\varepsilon>0} A(\varepsilon)\Big) \;\geq\; \wp\big(A(\varepsilon_0)\big)$$

for any choice of $\varepsilon_0 > 0$. Conversely, suppose $\wp\big(A(\varepsilon)\big) = 0$ for every positive ε. We claim that $X_n \to X$ a.s. To see this, observe that in the earlier characterization of the convergence of a sequence of real numbers, we could have restricted attention to values of ε of the form $\varepsilon = 1/k$ for positive integers k. In other words, a sequence of real numbers x_n converges to a real number x if and only if for every positive integer k, there is a positive integer N such that for all $n \geq N$, $|x_n \Leftrightarrow x| < 1/k$. Hence,

$$\wp(\{X_n \not\to X\}) \;=\; \wp\Big(\bigcup_{k=1}^{\infty} A(1/k)\Big) \;\leq\; \sum_{k=1}^{\infty} \wp\big(A(1/k)\big).$$

From this we see that if $\wp\big(A(\varepsilon)\big) = 0$ for all $\varepsilon > 0$, then $\wp(\{X_n \not\to X\}) = 0$.

To say more about almost sure convergence, we need to examine (11.6) more closely. Observe that

$$B_N(\varepsilon) \;=\; \bigcup_{n=N}^{\infty} A_n(\varepsilon) \;\supset\; \bigcup_{n=N+1}^{\infty} A_n(\varepsilon) \;=\; B_{N+1}(\varepsilon).$$

By limit property (1.5),

$$\wp\big(A(\varepsilon)\big) \;=\; \wp\Big(\bigcap_{N=1}^{\infty} B_N(\varepsilon)\Big) \;=\; \lim_{N\to\infty} \wp\big(B_N(\varepsilon)\big).$$

The next two examples use this equation to derive important results about almost sure convergence.

***Example* 11.10.** Show that if $X_n \to X$ a.s., then X_n converges in probability to X.

Solution. Recall that, by definition, X_n converges in probability to X if and only if $\wp\big(A_N(\varepsilon)\big) \to 0$ for every $\varepsilon > 0$. If $X_n \to X$ a.s., then for every $\varepsilon > 0$,

$$0 \;=\; \wp\big(A(\varepsilon)\big) \;=\; \lim_{N\to\infty} \wp\big(B_N(\varepsilon)\big).$$

Next, since

$$\wp\big(B_N(\varepsilon)\big) \;=\; \wp\Big(\bigcup_{n=N}^{\infty} A_n(\varepsilon)\Big) \;\geq\; \wp\big(A_N(\varepsilon)\big),$$

it follows that $\wp\big(A_N(\varepsilon)\big) \to 0$ too.

***Example* 11.11.** Show that if

$$\sum_{n=1}^{\infty} \wp\big(A_n(\varepsilon)\big) \;<\; \infty, \tag{11.7}$$

holds for all $\varepsilon > 0$, then $X_n \to X$ a.s.

Solution. For any $\varepsilon > 0$,

$$\begin{aligned}
\wp\big(A(\varepsilon)\big) &= \lim_{N\to\infty} \wp\big(B_N(\varepsilon)\big) \\
&= \lim_{N\to\infty} \wp\Big(\bigcup_{n=N}^{\infty} A_n(\varepsilon)\Big) \\
&\leq \lim_{N\to\infty} \sum_{n=N}^{\infty} \wp\big(A_n(\varepsilon)\big) \\
&= 0, \quad \text{on account of (11.7).}
\end{aligned}$$

What we have done here is derive a particular instance of the **Borel–Cantelli Lemma** (cf. Problem 15 in Chapter 1).

***Example* 11.12.** Let $X_1, X_2, \ldots$ be i.i.d. zero-mean random variables with finite fourth moment. Show that

$$M_n := \frac{1}{n}\sum_{i=1}^{n} X_i \;\to\; 0 \text{ a.s.}$$

Solution. We already know from Example 10.3 that M_n converges in mean square to zero, and hence, M_n converges to zero in probability and in distribution as well. Unfortunately, as shown in the next example, convergence in mean does not imply almost sure convergence. However, by the previous example, the almost sure convergence of M_n to zero will be established if we can show that for every $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \wp(|M_n| \geq \varepsilon) \ < \ \infty. \tag{11.8}$$

By Markov's inequality,

$$\wp(|M_n| \geq \varepsilon) \ = \ \wp(|M_n|^4 \geq \varepsilon^4) \ \leq \ \frac{\mathsf{E}[|M_n|^4]}{\varepsilon^4}.$$

By Problem 34, there are finite, nonnegative constants α (depending on $\mathsf{E}[X_i^4]$) and β (depending on $\mathsf{E}[X_i^2]$) such that

$$\mathsf{E}[|M_n|^4] \ \leq \ \frac{\alpha}{n^3} + \frac{\beta}{n^2}.$$

Hence, (11.8) holds by Problem 33.

The preceding example is an instance of the **strong law of large numbers**. The derivation in the example is quite simple because of the assumption of finite fourth moments (which implies finiteness of the third, second, and first moments by Lyapunov's inequality). A derivation assuming only finite second moments can be found in [17, pp. 326–327], and assuming only finite first moments in [17, pp. 329–331].

Strong Law of Large Numbers (SLLN). *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with finite mean m. Then*

$$M_n \ := \ \frac{1}{n}\sum_{i=1}^{n} X_i \ \to \ m \text{ a.s.}$$

Since almost sure convergence implies convergence in probability, the following form of the weak law of large numbers holds.

Weak Law of Large Numbers (WLLN). *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with finite mean m. Then*

$$M_n \ := \ \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in probability to m.

The weak law of large numbers in Example 11.1, which relied on the mean-square version in Example 10.3, required finite second moments and uncorrelated random variables. The above form does not require finite second moments, but does require independent random variables.

***Example* 11.13.** Let W_t be a Wiener process with $\mathsf{E}[W_t^2] = \sigma^2 t$. Use the strong law to show that W_n/n converges almost surely to zero.

Solution. Since $W_0 \equiv 0$, we can write

$$\frac{W_n}{n} \;=\; \frac{1}{n}\sum_{k=1}^{n}(W_k \Leftrightarrow W_{k-1}).$$

By the independent increments property of the Wiener process, the terms of the sum are i.i.d. $N(0,\sigma^2)$ random variables. By the strong law, this sum converges almost surely to zero.

***Example* 11.14.** We construct a sequence of random variables that converges in mean to zero, but does not converge almost surely to zero. Fix any positive integer n. Then n can be uniquely represented as $n = 2^m + k$, where m and k are integers satisfying $m \geq 0$ and $0 \leq k \leq 2^m \Leftrightarrow 1$. Define

$$g_n(x) \;=\; g_{2^m+k}(x) \;=\; I_{\left[\frac{k}{2^m},\,\frac{k+1}{2^m}\right)}(x).$$

For example, taking $m = 2$ and $k = 0, 1, 2, 3$, which corresponds to $n = 4, 5, 6, 7$, we find

$$g_4(x) \;=\; I_{\left[0,\,\frac{1}{4}\right)}(x), \quad g_5(x) \;=\; I_{\left[\frac{1}{4},\,\frac{1}{2}\right)}(x),$$

and

$$g_6(x) \;=\; I_{\left[\frac{1}{2},\,\frac{3}{4}\right)}(x), \quad g_7(x) \;=\; I_{\left[\frac{3}{4},\,1\right)}(x).$$

For fixed m, as k goes from 0 to $2^m \Leftrightarrow 1$, g_{2^m+k} is a sequence of pulses moving from left to right. This is repeated for $m+1$ with twice as many pulses that are half as wide. The two key ideas are that the pulses get narrower and that for any fixed $x \in [0,1)$, $g_n(x) = 1$ for infinitely many n.

Now let $U \sim$ uniform$[0,1)$. Then

$$\mathsf{E}[g_n(U)] \;=\; \wp\left(\frac{k}{2^m} \leq X < \frac{k+1}{2^m}\right) \;=\; \frac{1}{2^m}.$$

Since $m \to \infty$ as $n \to \infty$, we see that $g_n(U)$ converges in mean to zero. It then follows that $g_n(U)$ converges in probability to zero. Since almost sure convergence also implies convergence in probability, the only possible almost sure limit is zero.* However, we now show that $g_n(U)$ does not converge almost surely to zero. Fix any $x \in [0,1)$. Then for each $m = 0, 1, 2, \ldots$,

$$\frac{k}{2^m} \;\leq\; x \;<\; \frac{k+1}{2^m}$$

*Limits in probability are unique by Problem 4.

for some k satisfying $0 \le k \le 2^m \Leftrightarrow 1$. For these values of m and k, $g_{2^m+k}(x) = 1$. In other words, there are infinitely many values of $n = 2^m + k$ for which $g_n(x) = 1$. Hence, for $0 \le x < 1$, $g_n(x)$ does not converge to zero. Therefore,

$$\{U \in [0,1)\} \;\subset\; \{g_n(U) \not\to 0\},$$

and it follows that

$$\wp(\{g_n(U) \not\to 0\}) \;\ge\; \wp(\{U \in [0,1)\}) \;=\; 1.$$

The Skorohod Representation Theorem

The **Skorohod Representation Theorem** says that if X_n converges in distribution to X, then we can construct random variables Y_n and Y with $F_{Y_n} = F_{X_n}$, $F_Y = F_X$, and such that Y_n converges almost surely to Y. This can often simplify proofs concerning convergence in distribution.

***Example* 11.15.** Let X_n converge in distribution to X, and let $c(x)$ be a continuous function. Show that $c(X_n)$ converges in distribution to $c(X)$.

Solution. Let Y_n and Y be as given by the Skorohod representation theorem. Since Y_n converges almost surely to Y, the set

$$G \;:=\; \{\omega \in \Omega : Y_n(\omega) \to Y(\omega)\}$$

has the property that $\wp(G^c) = 0$. Fix any $\omega \in G$. Then $Y_n(\omega) \to Y(\omega)$. Since c is continuous,

$$c\big(Y_n(\omega)\big) \;\to\; c\big(Y(\omega)\big).$$

Thus, $c(Y_n)$ converges almost surely to $c(Y)$. Now recall that almost-sure convergence implies convergence in probability, which implies convergence in distribution. Hence, $c(Y_n)$ converges in distribution to $c(Y)$. To conclude, observe that since Y_n and X_n have the same cumulative distribution function, so do $c(Y_n)$ and $c(X_n)$. Similarly, $c(Y)$ and $c(X)$ have the same cumulative distribution function. Thus, $c(X_n)$ converges in distribution to $c(X)$.

We now derive the Skorohod representation theorem. For $0 < u < 1$, let

$$G_n(u) \;:=\; \inf\{x \in \mathbb{R} : F_{X_n}(x) \ge u\},$$

and

$$G(u) \;:=\; \inf\{x \in \mathbb{R} : F_X(x) \ge u\}.$$

By Problem 27(a) in Chapter 8,

$$G(u) \;\le\; x \;\Leftrightarrow\; u \;\le\; F_X(x),$$

or, equivalently,

$$G(u) \;>\; x \;\Leftrightarrow\; u \;>\; F_X(x),$$

and similarly for G_n and F_{X_n}. Now let $U \sim \text{uniform}(0,1)$. By Problem 27(b) in Chapter 8, $Y_n := G_n(U)$ and $Y := G(U)$ satisfy $F_{Y_n} = F_{X_n}$ and $F_Y = F_X$, respectively.

From the definition of G, it is easy to see that G is nondecreasing. Hence, its set of discontinuities, call it D, is at most countable (Problem 39), and so $\wp(U \in D) = 0$ by Problem 40. We show below that for $u \notin D$, $G_n(u) \to G(u)$. It then follows that

$$Y_n(\omega) := G_n\big(U(\omega)\big) \;\to\; G\big(U(\omega)\big) =: Y(\omega),$$

except for $\omega \in \{\omega : U(\omega) \in D\}$, which has probability zero.

Fix any $u \notin D$, and let $\varepsilon > 0$ be given. Then between $G(u) \Leftrightarrow \varepsilon$ and $G(u)$ we can select a point x,

$$G(u) \Leftrightarrow \varepsilon \;<\; x \;<\; G(u),$$

that is a continuity point of F_X. Since $x < G(u)$,

$$F_X(x) \;<\; u.$$

Since x is a continuity point of F_X, $F_{X_n}(x)$ must be close to $F_X(x)$ for large n. Thus, for large n, $F_{X_n}(x) < u$. But this implies $G_n(u) > x$. Thus,

$$G(u) \Leftrightarrow \varepsilon \;<\; x \;<\; G_n(u),$$

and it follows that

$$G(u) \;\le\; \varliminf_{n\to\infty} G_n(u).$$

To obtain the reverse inequality involving the $\varlimsup$, fix any u' with $u < u' < 1$. Fix any $\varepsilon > 0$, and select another continuity point of F_X, again called x, such that

$$G(u') \;<\; x \;<\; G(u') + \varepsilon.$$

Then $G(u') \le x$, and so $u' \le F_X(x)$. But then $u < F_X(x)$. Since $F_{X_n}(x) \to F_X(x)$, for large n, $u < F_{X_n}(x)$, which implies $G_n(u) \le x$. It then follows that

$$\varlimsup_{n\to\infty} G_n(u) \;\le\; G(u').$$

Since u is a continuity point of G, we can let $u' \to u$ to get

$$\varlimsup_{n\to\infty} G_n(u) \;\le\; G(u).$$

It now follows that $G_n(u) \to G(u)$ as claimed.

11.4. Notes

Notes §11.3: Almost Sure Convergence

***Note* 1.** In order that (11.3) be well defined, it is necessary that the set

$$\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}$$

be an **event** in the technical sense of Note 1 in Chapter 1. This is assured by the assumption that each X_n is a random variable (the term "random variable" is used in the technical sense of Note 1 in Chapter 2). The fact that this assumption is sufficient is demonstrated in more advanced texts, e.g., [4, pp. 183–184].

11.5. Problems

Problems §11.1: Convergence in Probability

1. Let c_n be a converging sequence of real numbers with limit c. Define the constant random variables $Y_n \equiv c_n$ and $Y \equiv c$. Show that Y_n converges in probability to Y.

2. Let $U \sim \text{uniform}[0,1]$, and put

$$X_n \;:=\; \sqrt{n} I_{[0,1/n]}(U), \quad n = 1, 2, \ldots.$$

   Does X_n converge in probability to zero?

3. Let V be any continuous random variable with an even density, and let c_n be any positive sequence with $c_n \to \infty$. Show that $X_n := V/c_n$ converges in probability to zero.

4. Show that limits in probability are unique; i.e., show that if X_n converges in probability to X, and X_n converges in probability to Y, then $\wp(X \neq Y) = 0$. *Hint:* Write

$$\{X \neq Y\} \;=\; \bigcup_{k=1}^{\infty} \{|X \Leftrightarrow Y| \geq 1/k\},$$

   and use limit property (1.4).

5. Suppose you have shown that given any $\varepsilon > 0$, for sufficiently large n,

$$\wp(|X_n \Leftrightarrow X| \geq \varepsilon) \;<\; \varepsilon.$$

   Show that

$$\lim_{n\to\infty} \wp(|X_n \Leftrightarrow X| \geq \varepsilon) \;=\; 0 \quad \text{for every } \varepsilon > 0.$$

6. Let $g(x,y)$ be continuous, and suppose that X_n converges in probability to X, and that Y_n converges in probability to Y. In this problem you will show that $g(X_n,Y_n)$ converges in probability to $g(X,Y)$.

   (a) Fix any $\varepsilon > 0$. Show that for sufficiently large α and β,

   $$\wp(|X| > \alpha) \;<\; \varepsilon/4 \quad \text{and} \quad \wp(|Y| > \beta) \;<\; \varepsilon/4.$$

   (b) Once α and β have been fixed, we can use the fact that $g(x,y)$ is uniformly continuous on the rectangle $|x| \le 2\alpha$ and $|y| \le 2\beta$. In other words, there is a $\delta > 0$ such that for all (x',y') and (x,y) in the rectangle and satisfying

   $$|x' \ominus x| \;\le\; \delta \quad \text{and} \quad |y' \ominus y| \;\le\; \delta,$$

   we have

   $$|g(x',y') \ominus g(x,y)| \;<\; \varepsilon.$$

   There is no loss of generality if we assume that $\delta \le \alpha$ and $\delta \le \beta$. Show that if the four conditions

   $$|X_n \ominus X| < \delta, \quad |Y_n \ominus Y| < \delta, \quad |X| \le \alpha, \quad \text{and} \quad |Y| \le \beta$$

   hold, then

   $$|g(X_n,Y_n) \ominus g(X,Y)| \;<\; \varepsilon.$$

   (c) Show that if n is large enough that

   $$\wp(|X_n \ominus X| \ge \delta) \;<\; \varepsilon/4 \quad \text{and} \quad \wp(|Y_n \ominus Y| \ge \delta) \;<\; \varepsilon/4,$$

   then

   $$\wp(|g(X_n,Y_n) \ominus g(X,Y)| \ge \varepsilon) \;<\; \varepsilon.$$

7. Let $X_1, X_2, \ldots$ be i.i.d. with common finite mean m and common finite variance σ^2. Also assume that $\mathsf{E}[X_i^4] < \infty$. Put

   $$M_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad V_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

   (a) Explain (briefly) why V_n converges in probability to $\sigma^2 + m^2$.

   (b) Explain (briefly) why

   $$S_n^2 \;:=\; \frac{n}{n \ominus 1}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) \ominus M_n^2\right]$$

   converges in probability to σ^2.

8. Let X_n converge in probability to X. Assume that there is a nonnegative random variable Y with $\mathsf{E}[Y] < \infty$ and such that $|X_n| \le Y$ for all n.

(a) Show that $\wp(|X| \geq Y + 1) = 0$. (In other words, $|X| < Y + 1$ with probability one, from which it follows that $\mathsf{E}[|X|] \leq \mathsf{E}[Y] + 1 < \infty$.)

(b) Show that X_n converges in mean to X. *Hints:* Write

$$\mathsf{E}[|X_n \Leftrightarrow X|] \;=\; \mathsf{E}[|X_n \Leftrightarrow X| I_{A_n}] + \mathsf{E}[|X_n \Leftrightarrow X| I_{A_n^c}],$$

where $A_n := \{|X_n \Leftrightarrow X| \geq \varepsilon\}$. Then use Problem 5 in Chapter 10 with $Z = Y + |X|$.

9. (a) Let g be a bounded, nonnegative function satisfying $\lim_{x \to 0} g(x) = 0$. Show that $\lim_{n \to \infty} \mathsf{E}[g(X_n)] = 0$ if X_n converges in probability to zero.

(b) Show that

$$\lim_{n \to \infty} \mathsf{E}\left[\frac{|X_n|}{1 + |X_n|}\right] \;=\; 0$$

if and only if X_n converges in probability to zero.

Problems §11.2: Convergence in Distribution

10. Let c_n be a converging sequence of real numbers with limit c. Define the constant random variables $Y_n \equiv c_n$ and $Y \equiv c$. Show that Y_n converges in distribution to Y.

11. Let X be a random variable, and let c_n be a positive sequence converging to limit c. Show that $c_n X$ converges in distribution to cX. Consider separately the cases $c = 0$ and $0 < c < \infty$.

12. For $t \geq 0$, let $X_t \leq Y_t \leq Z_t$ be three continuous-time random processes such that

$$\lim_{t \to \infty} F_{X_t}(y) \;=\; \lim_{t \to \infty} F_{Z_t}(y) \;=\; F(y)$$

for some continuous cdf F. Show that $F_{Y_t}(y) \to F(y)$ for all y.

13. Show that the Wiener integral $Y := \int_0^\infty g(\tau)\, dW_\tau$ is Gaussian with zero mean and variance $\int_0^\infty g(\tau)^2\, d\tau$. *Hints:* The desired integral Y is the mean-square limit of the sequence Y_n defined in Problem 18 in Chapter 10. Use Example 11.5.

14. Let $g(t, \tau)$ be such that for each t, $\int_0^\infty g(t, \tau)^2\, d\tau < \infty$. Define the process

$$X_t \;=\; \int_0^\infty g(t, \tau)\, dW_\tau.$$

Use the result of the preceding problem to show that for any $0 \leq t_1 < \cdots < t_n < \infty$, the random vector of samples $X := [X_{t_1}, \ldots, X_{t_n}]'$ is Gaussian. *Hint:* Read the first paragraph of Section 7.2.

15. If the moment generating functions $M_{X_n}(s)$ converge to the moment generating function $M_X(s)$, show that X_n converges in distribution to X. Also show that for nonnegative, integer-valued random variables, if the probability generating functions $G_{X_n}(z)$ converge to the probability generating function $G_X(z)$, then X_n converges in distribution to X. *Hint:* The Remark following Example 11.5 may be useful.

16. Let X_n and X be integer-valued random variables with probability mass functions $p_n(k) := \wp(X_n = k)$ and $p(k) := \wp(X = k)$, respectively.

    (a) If X_n converges in distribution to X, show that for each k, $p_n(k) \to p(k)$.

    (b) If X_n and X are nonnegative, and if for each $k \geq 0$, $p_n(k) \to p(k)$, show that X_n converges in distribution to X.

17. Let p_n be a sequence of numbers lying between 0 and 1 and such that $np_n \to \lambda > 0$ as $n \to \infty$. Let $X_n \sim \text{binomial}(n, p_n)$, and let $X \sim \text{Poisson}(\lambda)$. Show that X_n converges in distribution to X. *Hints:* By the previous problem, it suffices to prove that the probability mass functions converge. **Stirling's formula**,

$$n! \;\sim\; \sqrt{2\pi}\, n^{n+1/2} e^{-n},$$

    by which we mean

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi}\, n^{n+1/2} e^{-n}} \;=\; 1,$$

    and the formula

$$\lim_{n\to\infty} \left(1 \Leftrightarrow \frac{q_n}{n}\right)^n \;=\; e^{-q}, \quad \text{if } q_n \to q,$$

    may be helpful.

18. Let $X_n \sim \text{binomial}(n, p_n)$ and $X \sim \text{Poisson}(\lambda)$, where p_n and λ are as in the previous problem. Show that the probability generating function $G_{X_n}(z)$ converges to $G_X(z)$.

19. Suppose that X_n and X are such that for every bounded continuous function $g(x)$,

$$\lim_{n\to\infty} \mathsf{E}[g(X_n)] \;=\; \mathsf{E}[g(X)].$$

    Show that X_n converges in distribution to X as follows:

    (a) For $a < b$, sketch the three functions $I_{(-\infty,a]}(t)$, $I_{(-\infty,b]}(t)$, and

$$g_{a,b}(t) \;:=\; \begin{cases} 1, & t < a, \\ \dfrac{b \Leftrightarrow t}{b \Leftrightarrow a}, & a \leq t \leq b, \\ 0, & t > b. \end{cases}$$

(b) Your sketch in part (a) shows that

$$I_{(-\infty,a]}(t) \;\le\; g_{a,b}(t) \;\le\; I_{(-\infty,b]}(t).$$

Use these inequalities to show that for any random variable Y,

$$F_Y(a) \;\le\; \mathsf{E}[g_{a,b}(Y)] \;\le\; F_Y(b).$$

(c) For $\Delta x > 0$, use part (b) with $a = x$ and $b = x + \Delta x$ to show that

$$\overline{\lim_{n\to\infty}} F_{X_n}(x) \;\le\; F_X(x + \Delta x).$$

(d) For $\Delta x > 0$, use part (b) with $a = x \Leftrightarrow \Delta x$ and $b = x$ to show that

$$F_X(x \Leftrightarrow \Delta x) \;\le\; \underline{\lim_{n\to\infty}} F_{X_n}(x).$$

(e) If x is a continuity point of F_X, show that

$$\lim_{n\to\infty} F_{X_n}(x) \;=\; F_X(x).$$

20. Show that X_n converges in distribution to zero if and only if

$$\lim_{n\to\infty} \mathsf{E}\left[\frac{|X_n|}{1+|X_n|}\right] \;=\; 0.$$

21. For $t \ge 0$, let Z_t be a continuous-time random process. Suppose that as $t \to \infty$, $F_{Z_t}(z)$ converges to a continuous cdf $F(z)$. Let $u(t)$ be a positive function of t such that $u(t) \to 1$ as $t \to \infty$. Show that

$$\lim_{t\to\infty} \wp\big(Z_t \le z\, u(t)\big) \;=\; F(z).$$

Hint: Modify the derivation in Example 11.8.

22. Let Z_t be as in the preceding problem. Show that if $c(t) \to c > 0$, then

$$\lim_{t\to\infty} \wp\big(c(t)\, Z_t \le z\big) \;=\; F(c/z).$$

23. Let Z_t be as in Problem 21. Let $s(t) \to 0$ as $t \to \infty$. Show that if $X_t = Z_t + s(t)$, then $F_{X_t}(x) \to F(x)$.

24. Let N_t be a Poisson process of rate λ. Show that

$$Y_t \;:=\; \frac{\frac{N_t}{t} \Leftrightarrow \lambda}{\sqrt{\lambda/t}}$$

converges in distribution to an $N(0,1)$ random variable. *Hint:* By Example 11.7, Y_n converges in distribution to an $N(0,1)$ random variable. Next, since N_t is a nondecreasing function of t, observe that

$$N_{\lfloor t \rfloor} \;\le\; N_t \;\le\; N_{\lceil t \rceil},$$

where $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t, and $\lceil t \rceil$ denotes the smallest integer greater than or equal to t. *Hint:* The preceding two problems and Problem 12 may be useful.

Problems §11.3: Almost Sure Convergence

25. Let $X_n \to X$ a.s. and let $Y_n \to Y$ a.s. If $g(x,y)$ is a continuous function, show that $g(X_n, Y_n) \to g(X,Y)$ a.s.

26. Let $X_n \to X$ a.s., and suppose that $X = Y$ a.s. Show that $X_n \to Y$ a.s. (The statement $X = Y$ a.s. means $\wp(X \neq Y) = 0$.)

27. Show that almost sure limits are unique; i.e., if $X_n \to X$ a.s. and $X_n \to Y$ a.s., then $X = Y$ a.s. (The statement $X = Y$ a.s. means $\wp(X \neq Y) = 0$.)

28. Suppose $X_n \to X$ a.s. and $Y_n \to Y$ a.s. Show that if $X_n \leq Y_n$ a.s. for all n, then $X \leq Y$ a.s. (The statement $X_n \leq Y_n$ a.s. means $\wp(X_n > Y_n) = 0$.)

29. If X_n converges almost surely and in mean, show that the two limits are equal almost surely. *Hint:* Problem 4 may be helpful.

30. In Problem 11, suppose that the limit is $c = \infty$. Specify the limit random variable $cX(\omega)$ as a function of ω. Note that $cX(\omega)$ is a discrete random variable. What is its probability mass function?

31. Let S be a nonnegative random variable with $\mathsf{E}[S] < \infty$. Show that $S < \infty$ a.s. *Hints:* It is enough to show that $\wp(S = \infty) = 0$. Observe that

$$\{S = \infty\} \;=\; \bigcap_{n=1}^{\infty} \{S > n\}.$$

Now appeal to the limit property (1.5) and use Markov's inequality.

32. Under the assumptions of Problem 17 in Chapter 10, show that

$$\sum_{k=1}^{\infty} h_k X_k$$

is well defined as an almost-sure limit. *Hints:* It is enough to prove that

$$S \;:=\; \sum_{k=1}^{\infty} |h_k X_k| \;<\; \infty \quad \text{a.s.}$$

Hence, the result of the preceding problem can be applied if it can be shown that $\mathsf{E}[S] < \infty$. To this end, put

$$S_n \;:=\; \sum_{k=1}^{n} |h_k|\,|X_k|.$$

By Problem 17 in Chapter 10, S_n converges in mean to $S \in L^1$. Use the nonnegativity of S_n and S along with Problem 16 in Chapter 10 to show that

$$\mathsf{E}[S] \;=\; \lim_{n\to\infty} \sum_{k=1}^{n} |h_k| \mathsf{E}[|X_k|] \;<\; \infty.$$

33. For $p > 1$, show that $\sum_{n=1}^{\infty} 1/n^p < \infty$.

34. Let $X_1, X_2, \ldots$ be i.i.d. with $\gamma := \mathsf{E}[X_i^4]$, $\sigma^2 := \mathsf{E}[X_i^2]$, and $\mathsf{E}[X_i] = 0$. Show that

$$M_n \;:=\; \frac{1}{n} \sum_{i=1}^{n} X_i$$

satisfies $\mathsf{E}[M_n^4] = [n\gamma + 3n(n \Leftrightarrow 1)\sigma^4]/n^4$.

35. In Problem 7, explain why the assumption $\mathsf{E}[X_i^4] < \infty$ can be omitted.

36. Let N_t be a Poisson process of rate λ. Show that N_t/t converges almost surely to λ. *Hint:* First show that N_n/n converges almost surely to λ. Second, since N_t is a nondecreasing function of t, observe that

$$N_{\lfloor t \rfloor} \;\le\; N_t \;\le\; N_{\lceil t \rceil},$$

where $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t, and $\lceil t \rceil$ denotes the smallest integer greater than or equal to t.

37. Let N_t be a renewal process as defined in Section 8.2. Let $X_1, X_2, \ldots$ denote the i.i.d. interarrival times. Assume that the interarrival times have finite, positive mean μ.

    (a) Show that for any $\varepsilon > 0$, for all sufficiently large n,

$$\sum_{k=1}^{n} X_k \;<\; n(\mu + \varepsilon) \quad \text{a.s.}$$

    (b) Show that $N_t \to \infty$ as $t \to \infty$ a.s.; i.e., show that for any M, $N_t \ge M$ for all sufficiently large t.

    (c) Show that $n/T_n \to 1/\mu$ a.s. Here $T_n := X_1 + \cdots + X_n$ is the nth occurrence time.

    (d) Show that $N_t/t \to 1/\mu$ a.s. *Hints:* On account of (*c*), if we put $Y_n := n/T_n$, then $Y_{N_t} \to 1/\mu$ a.s. since $N_t \to \infty$. Also note that

$$T_{N_t} \;\le\; t \;<\; T_{N_t+1}.$$

38. Give an example of a sequence of random variables that converges almost surely to zero but not in mean.

39. Let G be a nondecreasing function defined on the closed interval $[a,b]$. Let D_ε denote the set of discontinuities of size greater than ε on $[a,b]$,

$$D_\varepsilon \;:=\; \{u \in [a,b] : G(u+) \Leftrightarrow G(u\Leftrightarrow) > \varepsilon\},$$

with the understanding that $G(b+)$ means $G(b)$ and $G(a\Leftrightarrow)$ means $G(a)$. Show that if there are n points in D_ε, then

$$n \;<\; [G(b) \Leftrightarrow G(a)]/\varepsilon + 2.$$

Remark. The set of all discontinuities of G on $[a,b]$, denoted by $D[a,b]$, is simply $\bigcup_{k=1}^{\infty} D_{1/k}$. Since this is a countable union of finite sets, $D[a,b]$ is at most countably infinite. If G is defined on the open interval $(0,1)$, we can write

$$D(0,1) \;=\; \bigcup_{n=3}^{\infty} D[1/n, 1 \Leftrightarrow 1/n].$$

Since this is a countable union of countably infinite sets, $D(0,1)$ is also countably infinite [37, p. 21, Proposition 7].

40. Let D be a countably infinite subset of $(0,1)$. Let $U \sim \text{uniform}(0,1)$. Show that $\wp(U \in D) = 0$. *Hint:* Since D is countably infinite, we can enumerate its elements as a sequence u_n. Fix any $\varepsilon > 0$ and put $K_n := (u_n \Leftrightarrow \varepsilon/2^n, u_n + \varepsilon/2^n)$. Observe that

$$D \;\subset\; \bigcup_{n=1}^{\infty} K_n.$$

Now bound

$$\wp\left(U \in \bigcup_{n=1}^{\infty} K_n\right).$$

CHAPTER 12

Parameter Estimation and Confidence Intervals

Consider observed measurements $X_1, X_2, \ldots$, each having unknown mean m. Then natural thing to do is to use the **sample mean**

$$M_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an estimate of m. However, since M_n is a random variable and m is a constant, we do not expect to have M_n always equal to m (or even ever equal if the X_i are continuous random variables). In this chapter we investigate how close the sample mean M_n is to the true mean m (also known as the **ensemble mean** or the **population mean**).

Section 12.1 introduces the notions of sample mean, unbiased estimator, and consistent estimator. It also recalls the central limit theorem approximation from Section 4.7. Section 12.2 introduces the notion of a confidence interval for the mean when the variance is known. In Section 12.3, the sample variance is introduced, and some of its properties presented. In Section 12.4, the sample mean and sample variance are combined to obtain confidence intervals for the mean when both the mean and the variance are unknown. Section 12.5 considers the special case of normal data. While the preceding sections assume that the number of samples is large, in the normal case, this assumption is not necessary. To obtain confidence intervals for the mean when the variance is unknown, we introduce the student's t density. By introducing the chi-squared density, we obtain confidence intervals for the variance when the mean is known and when the mean is unknown. Section 12.5 concludes with a subsection containing derivations of the more complicated results concerning the density of the sample variance and of the T random variable. These derivations can be skipped by the beginning student.

12.1. The Sample Mean

Although we do not expect to have $M_n = m$, we can say that

$$\mathsf{E}[M_n] = \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} m = m.$$

In other words, the expected value of the estimator is equal to the quantity we are trying to estimate. An estimator with this property is said to be **unbiased**.

Next, if the X_i have common finite variance σ^2, as we assume from now on, and if they are uncorrelated, then for any error bound $\delta > 0$, Chebyshev's

inequality tells us that

$$\wp(|M_n \Leftrightarrow m| > \delta) \;\leq\; \frac{\mathsf{E}[|M_n \Leftrightarrow m|^2]}{\delta^2} \;=\; \frac{\mathsf{var}(M_n)}{\delta^2} \;=\; \frac{\sigma^2}{n\delta^2}. \tag{12.1}$$

This says that the probability that the estimate M_n differs from the true mean m by more than δ is upper bounded by $\sigma^2/(n\delta^2)$. For large enough n, this bound is very small, and so the probability of being off by more than δ is negligible. In particular, (12.1) implies that

$$\lim_{n\to\infty} \wp(|M_n \Leftrightarrow m| > \delta) \;=\; 0, \quad \text{for every } \delta > 0.$$

The terminology for describing the above expression is to say that M_n **converges in probability** to m. An estimator that converges in probability to the parameter to be estimated is said to be **consistent**. Thus, the sample mean is a consistent estimator of the population mean.

While it is reassuring to know that the sample mean is a consistent estimator of the ensemble mean, this is an asymptotic result. In practice, we work with a finite value of n, and so it is unfortunate that, as the next example shows, the bound in (12.1) is not very good.

***Example* 12.1.** Let $X_1, X_2, \ldots$ be i.i.d. $N(m, \sigma^2)$. Then $M_n \sim N(m, \sigma^2/n)$ and so

$$Y_n \;:=\; \frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}}$$

is $N(0,1)$. Thus,

$$\wp(|M_n \Leftrightarrow m| > \sigma) \;=\; \wp(|Y_n| > \sqrt{n}\,) \;=\; 2[1 \Leftrightarrow \Phi(\sqrt{n}\,)],$$

where we have used the fact that Y_n has an even density. For $n = 3$, we find $\wp(|M_3 \Leftrightarrow m| > \sigma) = 0.0833$. For $\delta = \sigma$ in (12.1), we have the bound

$$\wp(|M_n \Leftrightarrow m| > \sigma) \;\leq\; \frac{1}{n}.$$

To make this bound less than or equal to 0.0833 would require $n \geq 12$. Thus, (12.1) would lead us to use a lot more data than is really necessary.

Since the bound in (12.1) is not very good, we would like to know the distribution of Y_n itself in the general case, not just the Gaussian case as in the example. Fortunately, the central limit theorem (Section 4.7) tells us that if the X_i are i.i.d. with common mean m and common variance σ^2, then for large n, F_{Y_n} can be approximated by the normal cdf Φ.

12.2. Confidence Intervals When the Variance Is Known

The notion of a confidence interval is obtained by rearranging

$$\wp\Big(|M_n \Leftrightarrow m| \le \frac{\sigma y}{\sqrt{n}}\Big)$$

as

$$\wp\Big(\Leftrightarrow\frac{\sigma y}{\sqrt{n}} \le M_n \Leftrightarrow m \le \frac{\sigma y}{\sqrt{n}}\Big),$$

or

$$\wp\Big(\frac{\sigma y}{\sqrt{n}} \ge m \Leftrightarrow M_n \ge \Leftrightarrow\frac{\sigma y}{\sqrt{n}}\Big),$$

and then

$$\wp\Big(M_n + \frac{\sigma y}{\sqrt{n}} \ge m \ge M_n \Leftrightarrow \frac{\sigma y}{\sqrt{n}}\Big).$$

The above quantity is the probability that the random interval

$$\Big[M_n \Leftrightarrow \frac{\sigma y}{\sqrt{n}}, M_n + \frac{\sigma y}{\sqrt{n}}\Big]$$

contains the true mean m. This random interval is called a **confidence interval**. If y is chosen so that

$$\wp\Big(m \in \Big[M_n \Leftrightarrow \frac{\sigma y}{\sqrt{n}}, M_n + \frac{\sigma y}{\sqrt{n}}\Big]\Big) \;=\; 1 \Leftrightarrow \alpha \tag{12.2}$$

for some $0 < \alpha < 1$, then the confidence interval is said to be a $100(1 \Leftrightarrow \alpha)\%$ confidence interval. The shorthand for the above expression is

$$m \;=\; M_n \pm \frac{\sigma y}{\sqrt{n}} \quad \text{with } 100(1 \Leftrightarrow \alpha)\% \text{ probability.}$$

In practice, we are given a **confidence level** $1 \Leftrightarrow \alpha$, and we would like to choose y so that (12.2) holds. Now, (12.2) is equivalent to

$$\begin{aligned}
1 \Leftrightarrow \alpha &= \wp\Big(|M_n \Leftrightarrow m| \le \frac{\sigma y}{\sqrt{n}}\Big) \\
&= \wp\Big(\frac{|M_n \Leftrightarrow m|}{\sigma/\sqrt{n}} \le y\Big) \\
&= \wp(|Y_n| \le y) \\
&= F_{Y_n}(y) \Leftrightarrow F_{Y_n}(\Leftrightarrow y) \\
&\approx \Phi(y) \Leftrightarrow \Phi(\Leftrightarrow y),
\end{aligned}$$

where the last step follows by the central limit theorem (Section 4.7) if the X_i are i.i.d. with common mean m and common variance σ^2. Since the $N(0,1)$ density is even, the approximation becomes

$$2\Phi(y) \Leftrightarrow 1 \;\approx\; 1 \Leftrightarrow \alpha.$$

$1 \Leftrightarrow \alpha$	$y_{\alpha/2}$
0.90	1.645
0.91	1.695
0.92	1.751
0.93	1.812
0.94	1.881
0.95	1.960
0.96	2.054
0.97	2.170
0.98	2.326
0.99	2.576

Table 12.1. Confidence levels $1-\alpha$ and corresponding $y_{\alpha/2}$ such that $2\Phi(y_{\alpha/2})-1 = 1-\alpha$.

Ignoring the approximation, we solve

$$\Phi(y) \;=\; 1 \Leftrightarrow \alpha/2$$

for y. We denote the solution of this equation by $y_{\alpha/2}$. It can be found from tables, e.g., Table 12.1, or numerically by finding the unique root of the equation $\Phi(y) + \alpha/2 \Leftrightarrow 1 = 0$, or in MATLAB by $y_{\alpha/2} =$ `norminv(1 ⇔alpha/2)`.*

The width of a confidence interval is

$$2 \cdot \frac{\sigma y_{\alpha/2}}{\sqrt{n}}.$$

Notice that as n increases, the width gets smaller as $1/\sqrt{n}$. If we increase the requested confidence level $1 \Leftrightarrow \alpha$, we see from Table 12.1 that $y_{\alpha/2}$ gets larger. Hence, by using a higher confidence level, the confidence interval gets larger.

***Example* 12.2.** Let $X_1, X_2, \ldots$ be i.i.d. random variables with variance $\sigma^2 = 2$. If $M_{100} = 7.129$, find the 93% and 97% confidence intervals for the population mean.

Solution. In Table 12.1 we scan the $1 \Leftrightarrow \alpha$ column until we find 0.93. The corresponding value of $y_{\alpha/2}$ is 1.812. Since $\sigma = \sqrt{2}$, we obtain the 93% confidence interval

$$\left[7.129 \Leftrightarrow \frac{1.812\sqrt{2}}{\sqrt{100}}, 7.129 + \frac{1.812\sqrt{2}}{\sqrt{100}}\right] \;=\; [6.873, 7.385].$$

Since $1.812\sqrt{2}/\sqrt{100} = 0.256$, we can also express this confidence interval as

$$m \;=\; 7.129 \pm 0.256 \text{ with } 93\% \text{ probability.}$$

*Since Φ can be related to the error function erf (see Section 4.1), $y_{\alpha/2}$ can also be found using the inverse of the error function. Hence, $y_{\alpha/2} = \sqrt{2}\,\mathrm{erf}^{-1}(1-\alpha)$. The MATLAB command for $\mathrm{erf}^{-1}(z)$ is `erfinv(z)`.

For the 97% confidence interval, we use $y_{\alpha/2} = 2.170$ and obtain

$$\left[7.129 \Leftrightarrow \frac{2.170\sqrt{2}}{\sqrt{100}}, 7.129 + \frac{2.170\sqrt{2}}{\sqrt{100}}\right] \;=\; [6.822, 7.436],$$

or

$$m \;=\; 7.129 \pm 0.307 \text{ with } 97\% \text{ probability.}$$

12.3. The Sample Variance

If the population variance σ^2 is unknown, then the confidence interval

$$\left[M_n \Leftrightarrow \frac{\sigma y}{\sqrt{n}}, M_n + \frac{\sigma y}{\sqrt{n}}\right]$$

cannot be used. Instead, we replace σ by the **sample standard deviation** S_n, where

$$S_n^2 \;:=\; \frac{1}{n \Leftrightarrow 1} \sum_{i=1}^{n} (X_i \Leftrightarrow M_n)^2$$

is the **sample variance**. This leads to the new confidence interval

$$\left[M_n \Leftrightarrow \frac{S_n y}{\sqrt{n}}, M_n + \frac{S_n y}{\sqrt{n}}\right].$$

Before we can do any analysis of this, we need some properties of S_n.

To begin, we need the formula,

$$S_n^2 \;=\; \frac{1}{n \Leftrightarrow 1}\left[\left(\sum_{i=1}^{n} X_i^2\right) \Leftrightarrow n M_n^2\right]. \tag{12.3}$$

It is derived as follows. Write

$$\begin{aligned}
S_n^2 \;&:=\; \frac{1}{n \Leftrightarrow 1} \sum_{i=1}^{n} (X_i \Leftrightarrow M_n)^2 \\
&=\; \frac{1}{n \Leftrightarrow 1} \sum_{i=1}^{n} (X_i^2 \Leftrightarrow 2 X_i M_n + M_n^2) \\
&=\; \frac{1}{n \Leftrightarrow 1}\left[\left(\sum_{i=1}^{n} X_i^2\right) \Leftrightarrow 2\left(\sum_{i=1}^{n} X_i\right) M_n + n M_n^2\right] \\
&=\; \frac{1}{n \Leftrightarrow 1}\left[\left(\sum_{i=1}^{n} X_i^2\right) \Leftrightarrow 2(n\, M_n) M_n + n M_n^2\right] \\
&=\; \frac{1}{n \Leftrightarrow 1}\left[\left(\sum_{i=1}^{n} X_i^2\right) \Leftrightarrow n M_n^2\right].
\end{aligned}$$

Using (12.3), it is shown in the problems that S_n^2 is an unbiased estimator of σ^2; i.e., $\mathsf{E}[S_n^2] = \sigma^2$. Furthermore, if the X_i are i.i.d. with finite fourth moment, then it is shown in the problems that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2$$

is a consistent estimator of the second moment $\mathsf{E}[X_i^2] = \sigma^2 + m^2$; i.e., the above formula converges in probability to $\sigma^2 + m^2$. Since M_n converges in probability to m, it follows that

$$S_n^2 \;=\; \frac{n}{n \Leftrightarrow 1}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) \Leftrightarrow M_n^2\right]$$

converges in probability[1] to $(\sigma^2 + m^2) \Leftrightarrow m^2 = \sigma^2$. In other words, S_n^2 is a consistent estimator of σ^2. Furthermore, it now follows that S_n is a consistent estimator of σ; i.e., S_n converges in probability to σ.

12.4. Confidence Intervals When the Variance Is Unknown

As noted in the previous section, if the population variance σ^2 is unknown, then the confidence interval

$$\left[M_n \Leftrightarrow \frac{\sigma y}{\sqrt{n}}, M_n + \frac{\sigma y}{\sqrt{n}}\right]$$

cannot be used. Instead, we replace σ by the sample standard deviation S_n. This leads to the new confidence interval

$$\left[M_n \Leftrightarrow \frac{S_n y}{\sqrt{n}}, M_n + \frac{S_n y}{\sqrt{n}}\right].$$

Given a confidence level $1 \Leftrightarrow \alpha$, we would like to choose y so that

$$\wp\left(m \in \left[M_n \Leftrightarrow \frac{S_n y}{\sqrt{n}}, M_n + \frac{S_n y}{\sqrt{n}}\right]\right) \;=\; 1 \Leftrightarrow \alpha. \tag{12.4}$$

To this end, rewrite the left-hand side as

$$\wp\left(|M_n \Leftrightarrow m| \le \frac{S_n\, y}{\sqrt{n}}\right),$$

or

$$\wp\left(\left|\frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}}\right| \le \frac{S_n\, y}{\sigma}\right).$$

Again using $Y_n := (M_n \Leftrightarrow m)/(\sigma/\sqrt{n}\,)$, we have

$$\wp\left(|Y_n| \le \frac{S_n\, y}{\sigma}\right).$$

As $n \to \infty$, S_n is converging in probability to σ. So,[2] for large n,

$$\begin{aligned} \wp\Big(|Y_n| \le \frac{S_n\, y}{\sigma}\Big) &\approx \wp(|Y_n| \le y) \\ &= F_{Y_n}(y) \Leftrightarrow F_{Y_n}(\Leftrightarrow y). \\ &\approx 2\Phi(y) \Leftrightarrow 1, \end{aligned} \tag{12.5}$$

where the last step follows by the central limit theorem approximation. Thus, if we want to solve (12.4), we can solve the approximate equation $2\Phi(y) \Leftrightarrow 1 = 1 \Leftrightarrow \alpha$ instead, exactly as in Section 12.2. In other words, to find confidence intervals, we still get $y_{\alpha/2}$ from Table 12.1, but we use S_n in place of σ.

Remark. In this section, we are using not only the central limit theorem approximation, for which it is suggested n should be at least 30, but we are also using the approximation (12.5). For the methods of this section to apply, n should be at least 100. This choice is motivated by considerations in the next section.

***Example* 12.3.** Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli(p) random variables. Find the 95% confidence interval for p if $M_{100} = 0.28$ and $S_{100} = 0.451$.

Solution. Observe that since $m := \mathsf{E}[X_i] = p$, we can use M_n to estimate p. From Table 12.1, $y_{\alpha/2} = 1.960$, $S_{100}\, y_{\alpha/2}/\sqrt{100} = 0.088$, and

$$p \;=\; 0.28 \pm 0.088 \text{ with } 95\% \text{ probability.}$$

The actual confidence interval is $[0.192, 0.368]$.

Applications

Estimating the number of defective products in a lot. Consider a production run of N cellular phones, of which, say d are defective. The only way to determine d exactly and for certain is to test every phone. This is not practical if N is very large. So we consider the following procedure to estimate the fraction of defectives, $p := d/N$, based on testing only n phones, where n is large, but smaller than N.

```
FOR i = 1 TO n
  Select a phone at random from the lot of N phones;
  IF the ith phone selected is defective
    LET X_i = 1;
  ELSE
    LET X_i = 0;
  END IF
  Return the phone to the lot;
END FOR
```

Because phones are returned to the lot (sampling with replacement), it is possible to test the same phone more than once. However, because the phones are always chosen from the same set of N phones, the X_i are i.i.d. with $\wp(X_i = 1) = d/N = p$. Hence, the central limit theorem applies, and we can use the method of Example 12.3 to estimate p and $d = Np$. For example, if $N = 1,000$ and we use the numbers from Example 12.3, we would estimate that

$$d = 280 \pm 88 \text{ with } 95\% \text{ probability.}$$

In other words, we are 95% sure that the number of defectives is between 192 and 368 for this particular lot of 1,000 phones.

If the phones were not returned to the lot after testing (sampling without replacement), the X_i would not be i.i.d. as required by the central limit theorem. However, in sampling with replacement when n is much smaller than N, the chances of testing the same phone twice are negligible. Hence, we can actually sample without replacement and proceed as above.

Predicting the Outcome of an Election. In order to predict the outcome of a presidential election, 4,000 registered voters are surveyed at random. 2,104 (more than half) say they will vote for candidate A, and the rest say they will vote for candidate B. To predict the outcome of the election, let p be the fraction of votes actually received by candidate A out of the total number of voters N (millions). Our poll samples $n = 4,000$, and $M_{4000} = 2,104/4,000 = 0.526$. Suppose that $S_{4000} = 0.499$. For a 95% confidence interval for p, $y_{\alpha/2} = 1.960$, $S_{4000}\, y_{\alpha/2}/\sqrt{4000} = 0.015$, and

$$p = 0.526 \pm 0.015 \text{ with } 95\% \text{ probability.}$$

Rounding off, we would predict that candidate A will receive 53% of the vote, with a margin of error of 2%. Thus, we are 95% sure that candidate A will win the election.

Sampling with and without Replacement

Consider sampling n items from a batch of N items, d of which are defective. If we sample with replacement, then the theory above worked out rather simply. We also argued briefly that if n is much smaller than N, then sampling without replacement would give essentially the same results. We now make this statement more precise.

To begin, recall that the central limit theorem says that for large n, $F_{Y_n}(y) \approx \Phi(y)$, where $Y_n = (M_n \Leftrightarrow m)/(\sigma/\sqrt{n})$, and $M_n = (1/n)\sum_{i=1}^{n} X_i$. If we sample with replacement and set $X_i = 1$ if the ith item is defective, then the X_i are i.i.d. Bernoulli(p) with $p = d/N$. When $X_1, X_2, \ldots$ are i.i.d. Bernoulli(p), we know that $\sum_{i=1}^{n} X_i$ is binomial(n, p) (e.g., by Example 2.20). Putting this all together, we obtain the **DeMoivre–Laplace Theorem**, which says that if $V \sim$ binomial(n, p) and n is large, then the cdf of $(V/n \Leftrightarrow p)/\sqrt{p(1 \Leftrightarrow p)/n}$ is approximately standard normal.

Now suppose we sample n items without replacement. Let U denote the number of defectives out of the n samples. It is shown in the Notes[3] that U has the **hypergeometric**(N, d, n) pmf,

$$\wp(U = k) \;=\; \frac{\binom{d}{k}\binom{N \Leftrightarrow d}{n \Leftrightarrow k}}{\binom{N}{n}}, \quad k = 0, \ldots, n.$$

As we show below, if n is much smaller than d, $N \Leftrightarrow d$, and N, then $\wp(U = k) \approx \wp(V = k)$. It then follows that the cdf of $(U/n \Leftrightarrow p)/\sqrt{p(1 \Leftrightarrow p)/n}$ is close to the cdf of $(V/n \Leftrightarrow p)/\sqrt{p(1 \Leftrightarrow p)/n}$, which is close to the standard normal cdf if n is large. (Thus, to make it all work we need n large, but still much smaller than d, $N \Leftrightarrow d$, and N.)

To show that $\wp(U = k) \approx \wp(V = k)$, write out $\wp(U = k)$ as

$$\frac{d!}{k!(d \Leftrightarrow k)!} \cdot \frac{(N \Leftrightarrow d)!}{(n \Leftrightarrow k)![(N \Leftrightarrow d) \Leftrightarrow (n \Leftrightarrow k)]!} \cdot \frac{n!(N \Leftrightarrow n)!}{N!}.$$

We can easily identify the factor $\binom{n}{k}$. Next, since $0 \le k \le n \ll d$,

$$\frac{d!}{(d \Leftrightarrow k)!} \;=\; d(d \Leftrightarrow 1) \cdots (d \Leftrightarrow k + 1) \;\approx\; d^k.$$

Similarly, since $0 \le k \le n \ll (N \Leftrightarrow d)$,

$$\frac{(N \Leftrightarrow d)!}{[(N \Leftrightarrow d) \Leftrightarrow (n \Leftrightarrow k)]!} \;=\; (N \Leftrightarrow d) \cdots [(N \Leftrightarrow d) \Leftrightarrow (n \Leftrightarrow k) + 1] \;\approx\; (N \Leftrightarrow d)^{n-k}.$$

Finally, since $n \ll N$,

$$\frac{(N \Leftrightarrow n)!}{N!} \;=\; \frac{1}{N(N \Leftrightarrow 1) \cdots (N \Leftrightarrow n + 1)} \;\approx\; \frac{1}{N^n}.$$

Writing $p = d/N$, we have

$$\wp(U = k) \;\approx\; \binom{n}{k} p^k (1 \Leftrightarrow p)^{n-k}.$$

12.5. Confidence Intervals for Normal Data

In this section we assume that $X_1, X_2, \ldots$ are i.i.d. $N(m, \sigma^2)$.

Estimating the Mean

Recall from Example 12.1 that $Y_n \sim N(0, 1)$. Hence, the analysis in Sections 12.1 and 12.2 shows that

$$\begin{aligned} \wp\Big(m \in \Big[M_n \Leftrightarrow \frac{\sigma y}{\sqrt{n}}, M_n + \frac{\sigma y}{\sqrt{n}}\Big]\Big) &= \wp(|Y_n| \le y) \\ &= 2\Phi(y) \Leftrightarrow 1. \end{aligned}$$

The point is that for normal data there is no central limit theorem approximation. Hence, we can determine confidence intervals as in Section 12.2 even if $n < 30$.

***Example* 12.4.** Let $X_1, X_2, \ldots$ be i.i.d. $N(m, 2)$. If $M_{10} = 5.287$, find the 90% confidence interval for m.

Solution. From Table 12.1 for $1 \ominus \alpha = 0.90$, $y_{\alpha/2}$ is 1.645. Since $\sigma = \sqrt{2}$, we obtain the 90% confidence interval

$$\left[5.287 \ominus \frac{1.645\sqrt{2}}{\sqrt{10}}, 5.287 + \frac{1.645\sqrt{2}}{\sqrt{10}}\right] \;=\; [4.551, 6.023].$$

Since $1.645\sqrt{2}/\sqrt{10} = 0.736$, we can also express this confidence interval as

$$m \;=\; 5.287 \pm 0.736 \text{ with } 90\% \text{ probability.}$$

Unfortunately, the results of Section 12.4 still involve approximation in (12.5) even if the X_i are normal. However, we now show how to compute the left-hand side of (12.4) exactly when the X_i are normal. The left-hand side of (12.4) is

$$\wp\left(|M_n \ominus m| \le \frac{S_n\, y}{\sqrt{n}}\right) \;=\; \wp\left(\left|\frac{M_n \ominus m}{S_n/\sqrt{n}}\right| \le y\right).$$

It is convenient to rewrite the above quotient as

$$T \;:=\; \frac{M_n \ominus m}{S_n/\sqrt{n}}.$$

It is shown later that if $X_1, X_2, \ldots$ are i.i.d. $N(m, \sigma^2)$, then T has a student's t density with $\nu = n \ominus 1$ degrees of freedom (defined in Problem 17 in Chapter 3). To compute $100(1 \ominus \alpha)\%$ confidence intervals, we must solve

$$\wp(|T| \le y) \;=\; 1 \ominus \alpha.$$

Since the density f_T is even, this is equivalent to

$$2F_T(y) \ominus 1 \;=\; 1 \ominus \alpha,$$

or $F_T(y) = 1 \ominus \alpha/2$. This can be solved using tables, e.g., Table 12.2, or numerically by finding the unique root of the equation $F_T(y) + \alpha/2 \ominus 1 = 0$, or in MATLAB by $y_{\alpha/2} = \texttt{tinv(1} \ominus \texttt{alpha/2, n} \ominus \texttt{1)}$.

***Example* 12.5.** Let $X_1, X_2, \ldots$ be i.i.d. $N(m, \sigma^2)$ random variables, and suppose $M_{10} = 5.287$. Further suppose that $S_{10} = 1.564$. Find the 90% confidence interval for m.

Solution. In Table 12.2 with $n = 10$, we see that for $1 \ominus \alpha = 0.90$, $y_{\alpha/2}$ is 1.833. Since $S_{10}\, y_{\alpha/2}/\sqrt{10} = 0.907$,

$$m \;=\; 5.287 \pm 0.907 \text{ with } 90\% \text{ probability.}$$

The corresponding confidence interval is $[4.380, 6.194]$.

$1 \Leftrightarrow \alpha$	$y_{\alpha/2}(n = 10)$
0.90	1.833
0.91	1.899
0.92	1.973
0.93	2.055
0.94	2.150
0.95	2.262
0.96	2.398
0.97	2.574
0.98	2.821
0.99	3.250

$1 \Leftrightarrow \alpha$	$y_{\alpha/2}(n = 100)$
0.90	1.660
0.91	1.712
0.92	1.769
0.93	1.832
0.94	1.903
0.95	1.984
0.96	2.081
0.97	2.202
0.98	2.365
0.99	2.626

Table 12.2. Confidence levels $1-\alpha$ and corresponding $y_{\alpha/2}$ such that $\wp(|T| \leq y_{\alpha/2}) = 1-\alpha$. The left-hand table is for $n = 10$ observations with T having $n - 1 = 9$ degrees of freedom, and the right-hand table is for $n = 100$ observations with T having $n - 1 = 99$ degrees of freedom.

Limiting t Distribution

If we compare the $n = 100$ table in Table 12.2 with Table 12.1, we see they are almost the same. This is a consequence of the fact that as n increases, the t cdf converges to the standard normal cdf. We can see this by writing

$$\begin{aligned}
\wp(T \leq t) &= \wp\left(\frac{M_n \Leftrightarrow m}{S_n/\sqrt{n}} \leq t\right) \\
&= \wp\left(\frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}} \leq \frac{S_n}{\sigma} t\right) \\
&= \wp\left(Y_n \leq \frac{S_n}{\sigma} t\right) \\
&\approx \wp(Y_n \leq t),
\end{aligned}$$

since[2] S_n converges in probability to σ. Finally, since the X_i are independent and normal, $F_{Y_n}(t) = \Phi(t)$.

We also recall from Problem 18 in Chapter 3 that the t density converges to the standard normal density.

Estimating the Variance — Known Mean

Suppose that $X_1, X_2, \ldots$ are i.i.d. $N(m, \sigma^2)$ with m known but σ^2 unknown. We use

$$V_n^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i \Leftrightarrow m)^2$$

as our estimator of the variance σ^2. It is shown in the problems that V_n^2 is a consistent estimator of σ^2.

For determining confidence intervals, it is easier to work with

$$\frac{n}{\sigma^2}V_n^2 \;=\; \sum_{i=1}^{n}\left(\frac{X_i \Leftrightarrow m}{\sigma}\right)^2.$$

Since $(X_i \Leftrightarrow m)/\sigma$ is $N(0,1)$, its square is chi-squared with one degree of freedom (Problem 41 in Chapter 3 or Problem 7 in Chapter 4). It then follows that $\frac{n}{\sigma^2}V_n^2$ is chi-squared with n degrees of freedom (see Problem 46(c) and its Remark in Chapter 3).

Choose $0 < \ell < u$, and consider the equation

$$\wp\Big(\ell \le \frac{n}{\sigma^2}V_n^2 \le u\Big) \;=\; 1 \Leftrightarrow \alpha.$$

We can rewrite this as

$$\wp\Big(\frac{nV_n^2}{\ell} \ge \sigma^2 \ge \frac{nV_n^2}{u}\Big) \;=\; 1 \Leftrightarrow \alpha.$$

This suggests the confidence interval

$$\Big[\frac{nV_n^2}{u}, \frac{nV_n^2}{\ell}\Big].$$

Then the probability that σ^2 lies in this interval is

$$F(u) \Leftrightarrow F(\ell) \;=\; 1 \Leftrightarrow \alpha,$$

where F is the chi-squared cdf with n degrees of freedom. We usually choose ℓ and u to solve

$$F(\ell) \;=\; \alpha/2 \quad \text{and} \quad F(u) = 1 \Leftrightarrow \alpha/2.$$

These equations can be solved using tables, e.g., Table 12.3, or numerically by root finding, or in MATLAB with the commands $\ell =$ `chi2inv(alpha/2,n)` and $u =$ `chi2inv(1` $\Leftrightarrow$ `alpha/2,n)`.

***Example* 12.6.** Let $X_1, X_2, \ldots$ be i.i.d. $N(5, \sigma^2)$ random variables. Suppose that $V_{100}^2 = 1.645$. Find the 90% confidence interval for σ^2.

Solution. From Table 12.3 we see that for $1 \Leftrightarrow \alpha = 0.90$, $\ell = 77.929$ and $u = 124.342$. The 90% confidence interval is

$$\Big[\frac{100(1.645)}{124.342}, \frac{100(1.645)}{77.929}\Big] \;=\; [1.323, 2.111].$$

$1 \Leftrightarrow \alpha$	ℓ	u
0.90	77.929	124.342
0.91	77.326	125.170
0.92	76.671	126.079
0.93	75.949	127.092
0.94	75.142	128.237
0.95	74.222	129.561
0.96	73.142	131.142
0.97	71.818	133.120
0.98	70.065	135.807
0.99	67.328	140.169

Table 12.3. Confidence levels $1 - \alpha$ and corresponding values of ℓ and u such that $\wp(\ell \leq \frac{n}{\sigma^2} V_n^2 \leq u) = 1 - \alpha$ and such that $\wp(\frac{n}{\sigma^2} V_n^2 \leq \ell) = \wp(\frac{n}{\sigma^2} V_n^2 \geq u) = \alpha/2$ for $n = 100$ observations.

Estimating the Variance — Unknown Mean

Let $X_1, X_2, \ldots$ be i.i.d. $N(m, \sigma^2)$, where both the mean and the variance are unknown, but we are interested only in estimating the variance. Since we do not know m, we cannot use the estimator V_n^2 above. Instead we use S_n^2. However, for determining confidence intervals, it is easier to work with $\frac{n-1}{\sigma^2} S_n^2$. As argued below, $\frac{n-1}{\sigma^2} S_n^2$ is a chi-squared random variable with $n \Leftrightarrow 1$ degrees of freedom.

Choose $0 < \ell < u$, and consider the equation

$$\wp\Big(\ell \leq \frac{n \Leftrightarrow 1}{\sigma^2} S_n^2 \leq u\Big) \;=\; 1 \Leftrightarrow \alpha\,.$$

We can rewrite this as

$$\wp\Big(\frac{(n \Leftrightarrow 1) S_n^2}{\ell} \geq \sigma^2 \geq \frac{(n \Leftrightarrow 1) S_n^2}{u}\Big) \;=\; 1 \Leftrightarrow \alpha\,.$$

This suggests the confidence interval

$$\Big[\frac{(n \Leftrightarrow 1) S_n^2}{u}, \frac{(n \Leftrightarrow 1) S_n^2}{\ell}\Big].$$

Then the probability that σ^2 lies in this interval is

$$F(u) \Leftrightarrow F(\ell) \;=\; 1 \Leftrightarrow \alpha,$$

where now F is the chi-squared cdf with $n \Leftrightarrow 1$ degrees of freedom. We usually choose ℓ and u to solve

$$F(\ell) \;=\; \alpha/2 \quad \text{and} \quad F(u) = 1 \Leftrightarrow \alpha/2.$$

These equations can be solved using tables, e.g., Table 12.4, or numerically by root finding, or in MATLAB with the commands $\ell =$ `chi2inv(alpha/2, n⇔1)` and $u =$ `chi2inv(1⇔alpha/2, n⇔1)`.

$1 \Leftrightarrow \alpha$	ℓ	u
0.90	77.046	123.225
0.91	76.447	124.049
0.92	75.795	124.955
0.93	75.077	125.963
0.94	74.275	127.103
0.95	73.361	128.422
0.96	72.288	129.996
0.97	70.972	131.966
0.98	69.230	134.642
0.99	66.510	138.987

Table 12.4. Confidence levels $1-\alpha$ and corresponding values of ℓ and u such that $\wp(\ell \le \frac{n-1}{\sigma^2}S_n^2 \le u) = 1-\alpha$ and such that $\wp(\frac{n-1}{\sigma^2}S_n^2 \le \ell) = \wp(\frac{n-1}{\sigma^2}S_n^2 \ge u) = \alpha/2$ for $n = 100$ observations ($n-1 = 99$ degrees of freedom).

***Example* 12.7.** Let $X_1, X_2, \ldots$ be i.i.d. $N(m, \sigma^2)$ random variables. If $S_{100}^2 = 1.608$, find the 90% confidence interval for σ^2.

Solution. From Table 12.4 we see that for $1 \Leftrightarrow \alpha = 0.90$, $\ell = 77.046$ and $u = 123.225$. The 90% confidence interval is

$$\left[\frac{99(1.608)}{123.225}, \frac{99(1.608)}{77.046}\right] \;=\; [1.292, 2.067].$$

⋆ *Derivations*

The remainder of this section is devoted to deriving the distributions of S_n^2 and

$$T \;:=\; \frac{M_n \Leftrightarrow m}{S_n/\sqrt{n}} \;=\; \frac{(M_n \Leftrightarrow m)/(\sigma/\sqrt{n})}{\sqrt{\frac{n-1}{\sigma^2}S_n^2/(n \Leftrightarrow 1)}}$$

under the assumption that the X_i are i.i.d. $N(m, \sigma^2)$. The analysis is rather complicated, and may be omitted by the beginning student.

We begin with the numerator in T. Recall that $(M_n \Leftrightarrow m)/(\sigma/\sqrt{n}\,)$ is simply Y_n as defined in Section 12.1. It was shown in Example 12.1 that $Y_n \sim N(0,1)$.

For the denominator, we show that the density of $\frac{n-1}{\sigma^2}S_n^2$ is chi-squared with $n \Leftrightarrow 1$ degrees of freedom. We begin by recalling the derivation of (12.3). If we replace the first line of the derivation with

$$S_n^2 \;=\; \frac{1}{n \Leftrightarrow 1}\sum_{i=1}^{n}([X_i \Leftrightarrow m] \Leftrightarrow [M_n \Leftrightarrow m])^2,$$

then we end up with

$$S_n^2 \;=\; \frac{1}{n \Leftrightarrow 1}\left[\left(\sum_{i=1}^{n}[X_i \Leftrightarrow m]^2\right) \Leftrightarrow n[M_n \Leftrightarrow m]^2\right].$$

Using the notation $Z_i := (X_i \Leftrightarrow m)/\sigma$, we have

$$\frac{n \Leftrightarrow 1}{\sigma^2} S_n^2 \;=\; \sum_{i=1}^{n} Z_i^2 \Leftrightarrow n \left(\frac{M_n \Leftrightarrow m}{\sigma} \right)^2,$$

or

$$\frac{n \Leftrightarrow 1}{\sigma^2} S_n^2 + \left(\frac{M_n \Leftrightarrow m}{\sigma/\sqrt{n}} \right)^2 \;=\; \sum_{i=1}^{n} Z_i^2 .$$

As we argue below, the two terms on the left are independent. It then follows that the density of $\sum_{i=1}^{n} Z_i^2$ is equal to the convolution of the densities of the other two terms. To find the density of $\frac{n-1}{\sigma^2} S_n^2$, we use moment generating functions. Now, the second term on the left is the square of an $N(0,1)$ random variable, and by Problem 7 in Chapter 4 has a chi-squared density with one degree of freedom. Its moment generating function is $1/(1 \Leftrightarrow 2s)^{1/2}$ (see Problem 40(c) in Chapter 3). For the same reason, each Z_i^2 has a chi-squared density with one degree of freedom. Since the Z_i are independent, $\sum_{i=1}^{n} Z_i^2$ has a chi-squared density with n degrees of freedom, and its moment generating function is $1/(1 \Leftrightarrow 2s)^{n/2}$ (see Problem 46(c) in Chapter 3). It now follows that the moment generating function of $\frac{n-1}{\sigma^2} S_n^2$ is the quotient

$$\frac{1/(1 \Leftrightarrow 2s)^{n/2}}{1/(1 \Leftrightarrow 2s)^{1/2}} \;=\; \frac{1}{(1 \Leftrightarrow 2s)^{(n-1)/2}},$$

which is the moment generating function of a chi-squared density with $n \Leftrightarrow 1$ degrees of freedom.

It remains to show that S_n^2 and M_n are independent. Observe that S_n^2 is a function of the vector

$$W \;:=\; [(X_1 \Leftrightarrow M_n), \ldots, (X_n \Leftrightarrow M_n)]'.$$

In fact, $S_n^2 = W'W/(n \Leftrightarrow 1)$. By Example 7.5, the vector W and the sample mean are independent. It then follows that any function of W and any function of M_n are independent.

We can now find the density of

$$T \;=\; \frac{(M_n \Leftrightarrow m)/(\sigma/\sqrt{n})}{\sqrt{\frac{n-1}{\sigma^2} S_n^2/(n \Leftrightarrow 1)}}.$$

If the X_i are i.i.d. $N(m, \sigma^2)$, then the numerator and the denominator are independent; the numerator is $N(0,1)$, and the denominator is chi-squared with $n \Leftrightarrow 1$ degrees of freedom divided by $n \Leftrightarrow 1$. By Problem 24 in Chapter 5, T has a student's t density with $\nu = n \Leftrightarrow 1$ degrees of freedom.

12.6. Notes

Notes §12.3: The Sample Variance

***Note* 1.** We are appealing to the fact that if $g(u,v)$ is a continuous function of two variables, and if U_n converges in probability to u, and if V_n converges in probability to v, then $g(U_n, V_n)$ converges in probability to $g(u,v)$. This result is proved in Example 11.2.

Notes §12.4: Confidence Intervals When the Variance Is Unknown

***Note* 2.** We are appealing to the fact that if the cdf of Y_n, say F_n, converges to a continuous cdf F, and if U_n converges in probability to 1, then

$$\wp(Y_n \le yU_n) \;\to\; F(y).$$

This result, which is proved in Example 11.8, is a version of **Slutsky's Theorem**.

***Note* 3.** The hypergeometric random variable arises in the following situation. We have a collection of N items, d of which are defective. Rather than test all N items, we select at random a small number of items, say $n < N$. Let Y_n denote the number of defectives out the n items tested. We show that

$$\wp(Y_n = k) \;=\; \frac{\binom{d}{k}\binom{N \Leftrightarrow d}{n \Leftrightarrow k}}{\binom{N}{n}}, \quad k = 0, \ldots, n.$$

We denote this by $Y_n \sim \text{hypergeometric}(N, d, n)$.

Remark. In the typical case, $d \ge n$ and $N \Leftrightarrow d \ge n$; however, if these conditions do not hold in the above formula, it is understood that $\binom{d}{k} = 0$ if $d < k \le n$, and $\binom{N-d}{n-k} = 0$ if $n \Leftrightarrow k > N \Leftrightarrow d$, i.e., if $0 \le k < n \Leftrightarrow (N \Leftrightarrow d)$.

For $i = 1, \ldots, n$, draw at random an item from the collection and test it. If the ith item is defective, let $X_i = 1$, and put $X_i = 0$ otherwise. In either case, do *not* put the tested item back into the collection (sampling without replacement). Then the total number of defectives among the first n items tested is

$$Y_n \;:=\; \sum_{i=1}^{n} X_i.$$

We show that $Y_n \sim \text{hypergeometric}(N, d, n)$.

Consider the case $n = 1$. Then $Y_1 = X_1$, and the chance of drawing a defective item at random is simply the ratio of the number of defectives to the total number of items in the collection; i.e., $\wp(Y_1 = 1) = \wp(X_1 = 1) = d/N$.

Now in general, suppose the result is true for some $n \geq 1$. We show it is true for $n+1$. Use the law of total probability to write

$$\wp(Y_{n+1} = k) \;=\; \sum_{i=0}^{n} \wp(Y_{n+1} = k | Y_n = i) \wp(Y_n = i). \tag{12.6}$$

Since $Y_{n+1} = Y_n + X_{n+1}$, we can use the substitution law to write

$$\begin{aligned} \wp(Y_{n+1} = k | Y_n = i) &= \wp(Y_n + X_{n+1} = k | Y_n = i) \\ &= \wp(i + X_{n+1} = k | Y_n = i) \\ &= \wp(X_{n+1} = k \Leftrightarrow i | Y_n = i). \end{aligned}$$

Since X_{n+1} takes only the values zero and one, this last expression is zero unless $i = k$ or $i = k \Leftrightarrow 1$. Returning to (12.6), we can write

$$\wp(Y_{n+1} = k) \;=\; \sum_{i=k-1}^{k} \wp(X_{n+1} = k \Leftrightarrow i | Y_n = i) \wp(Y_n = i). \tag{12.7}$$

When $i = k \Leftrightarrow 1$, the above conditional probability is

$$\wp(X_{n+1} = 1 | Y_n = k \Leftrightarrow 1) \;=\; \frac{d \Leftrightarrow (k \Leftrightarrow 1)}{N \Leftrightarrow n},$$

since given $Y_n = k \Leftrightarrow 1$, there are $N \Leftrightarrow n$ items left in the collection, and of those, the number of defectives remaining is $d \Leftrightarrow (k \Leftrightarrow 1)$. When $i = k$, the needed conditional probability is

$$\wp(X_{n+1} = 0 | Y_n = k) \;=\; \frac{(N \Leftrightarrow d) \Leftrightarrow (n \Leftrightarrow k)}{N \Leftrightarrow n},$$

since given $Y_n = k$, there are $N \Leftrightarrow n$ items left in the collection, and of those, the number of *non*defectives remaining is $(N \Leftrightarrow d) \Leftrightarrow (n \Leftrightarrow k)$. If we now assume that $Y_n \sim \text{hypergeometric}(N, d, n)$, we can expand (12.7) to get

$$\begin{aligned} \wp(Y_{n+1} = k) \;=\; & \frac{d \Leftrightarrow (k \Leftrightarrow 1)}{N \Leftrightarrow n} \cdot \frac{\binom{d}{k \Leftrightarrow 1} \binom{N \Leftrightarrow d}{n \Leftrightarrow (k \Leftrightarrow 1)}}{\binom{N}{n}} \\ & + \frac{(N \Leftrightarrow d) \Leftrightarrow (n \Leftrightarrow k)}{N \Leftrightarrow n} \cdot \frac{\binom{d}{k} \binom{N \Leftrightarrow d}{n \Leftrightarrow k}}{\binom{N}{n}}. \end{aligned}$$

It is a simple calculation to see that the first term on the right is equal to

$$\left(1 \Leftrightarrow \frac{k}{n+1}\right) \cdot \frac{\binom{d}{k} \binom{N \Leftrightarrow d}{[n+1] \Leftrightarrow k}}{\binom{N}{n+1}},$$

and the second term is equal to

$$\frac{k}{n+1} \cdot \frac{\binom{d}{k}\binom{N \Leftrightarrow d}{[n+1] \Leftrightarrow k}}{\binom{N}{n+1}}.$$

Thus, $Y_{n+1} \sim \text{hypergeometric}(N, d, n+1)$.

12.7. Problems

Problems §12.1: The Sample Mean

1. Show that the random variable Y_n defined in Example 12.1 is $N(0,1)$.

2. Show that when the X_i are uncorrelated, Y_n has zero mean and unit variance.

Problems §12.2: Confidence Intervals When the Variance Is Known

3. If $\sigma^2 = 4$ and $n = 100$, how wide is the 99% confidence interval? How large would n have to be to have a 99% confidence interval of width less than or equal to $1/4$?

4. Let $W_1, W_2, \ldots$ be i.i.d. with zero mean and variance 4. Let $X_i = m + W_i$, where m is an unknown constant. If $M_{100} = 14.846$, find the 95% confidence interval.

5. Let $X_i = m + W_i$, where m is an unknown constant, and the W_i are i.i.d. Cauchy with parameter 1. Find $\delta > 0$ such that the probability is $2/3$ that the confidence interval $[M_n \Leftrightarrow \delta, M_n + \delta]$ contains m; i.e., find $\delta > 0$ such that

$$\wp(|M_n \Leftrightarrow m| \leq \delta) \;=\; 2/3.$$

Hints: Since $\mathsf{E}[W_i^2] = \infty$, the central limit theorem does not apply. However, you can solve for δ exactly if you can find the cdf of $M_n \Leftrightarrow m$. The cdf of W_i is $F(w) = \frac{1}{\pi}\tan^{-1}(w) + 1/2$, and the characteristic function of W_i is $\mathsf{E}[e^{j\nu W_i}] = e^{-|\nu|}$.

Problems §12.3: The Sample Variance

6. If $X_1, X_2, \ldots$ are uncorrelated, use the formula

$$S_n^2 \;=\; \frac{1}{n \Leftrightarrow 1}\left[\left(\sum_{i=1}^{n} X_i^2\right) \Leftrightarrow n M_n^2\right]$$

to show that S_n^2 is an unbiased estimator of σ^2; i.e., show that $\mathsf{E}[S_n^2] = \sigma^2$.

7. Let $X_1, X_2, \ldots$ be i.i.d. with finite fourth moment. Let $m_2 := \mathsf{E}[X_i^2]$ be the second moment. Show that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2$$

is a consistent estimator of m_2. *Hint:* Let $\tilde{X}_i := X_i^2$.

Problems §12.4: Confidence Intervals When the Variance is Unknown

8. Let $X_1, X_2, \ldots$ be i.i.d. random variables with unknown, finite mean m and variance σ^2. If $M_{100} = 10.083$ and $S_{100} = 0.568$, find the 95% confidence interval for the population mean.

9. Suppose that 100 engineering freshmen are selected at random and $X_1, \ldots, X_{100}$ are their times (in years) to graduation. If $M_{100} = 4.422$ and $S_{100} = 0.957$, find the 93% confidence interval for their expected time to graduate.

10. From a batch of $N = 10,000$ computers, $n = 100$ are sampled, and 10 are found defective. Estimate the number of defective computers in the total batch of $10,000$, and give the margin of error for 90% probability if $S_{100} = 0.302$.

11. You conduct a presidential preference poll by surveying $3,000$ voters. You find that $1,559$ (more than half) say they plan to vote for candidate A, and the others say they plan to vote for candidate B. If $S_{3000} = 0.500$, are you 90% sure that candidate A will win the election? Are you 99% sure?

12. From a batch of $100,000$ airbags, 500 are sampled, and 48 are found defective. Estimate the number of defective airbags in the total batch of $100,000$, and give the margin of error for 94% probability if $S_{100} = 0.295$.

13. A new vaccine has just been developed at your company. You need to be 97% sure that side effects do not occur more than 10% of the time.

    (a) In order to estimate the probability p of side effects, the vaccine is tested on 100 volunteers. Side effects are experienced by 6 of the volunteers. Using the value $S_{100} = 0.239$, find the 97% confidence interval for p if $S_{100} = 0.239$. Are you 97% sure that $p \leq 0.1$?

    (b) Another study is performed, this time with 1000 volunteers. Side effects occur in 71 volunteers. Find the 97% confidence interval for the probability p of side effects if $S_{1000} = 0.257$. Are you 97% sure that $p \leq 0.1$?

14. Packet transmission times on a certain Internet link are independent and identically distributed. Assume that the times have an exponential density with mean μ.

(a) Find the probability that in transmitting n packets, at least one of them takes more than t seconds to transmit.

(b) Let T denote the total time to transmit n packets. Find a closed-form expression for the density of T.

(c) Your answers to parts (a) and (b) depend on μ, which in practice is unknown and must be estimated. To estimate the expected transmission time, $n = 100$ packets are sent, and the transmission times $T_1, \ldots, T_n$ recorded. It is found that the sample mean $M_{100} = 1.994$, and sample standard deviation $S_{100} = 1.798$, where

$$M_n := \frac{1}{n}\sum_{i=1}^{n} T_i \quad \text{and} \quad S_n^2 := \frac{1}{n \hat{\Leftrightarrow} 1}\sum_{i=1}^{n}(T_i \hat{\Leftrightarrow} M_n)^2.$$

Find the 95% confidence interval for the expected transmission time.

15. Let N_t be a Poisson process with unknown intensity λ. For $i = 1, \ldots, 100$, put $X_i = (N_i \hat{\Leftrightarrow} N_{i-1})$. Then the X_i are i.i.d. Poisson(λ), and $\mathsf{E}[X_i] = \lambda$. If $M_{100} = 5.170$, and $S_{100} = 2.132$, find the 95% confidence interval for λ.

Problems §12.5: Confidence Intervals for Normal Data

16. Let $W_1, W_2, \ldots$ be i.i.d. $N(0, \sigma^2)$ with σ^2 unknown. Let $X_i = m + W_i$, where m is an unknown constant. Suppose $M_{10} = 14.832$ and $S_{10} = 1.904$. Find the 95% confidence interval for m.

17. Let $X_1, X_2, \ldots$ be i.i.d. $N(0, \sigma^2)$ with σ^2 unknown. Find the 95% confidence interval for σ^2 if $V_{100}^2 = 4.413$.

18. Let W_t be a zero-mean, wide-sense stationary white noise process with power spectral density $S_W(f) = \mathcal{N}_0/2$. Suppose that W_t is applied to the ideal lowpass filter of bandwidth $B = 1$ MHz and power gain 120 dB; i.e., $H(f) = GI_{[-B,B]}(f)$, where $G = 10^6$. Denote the filter output by Y_t, and for $i = 1, \ldots, 100$, put $X_i := Y_{i\Delta t}$, where $\Delta t = (2B)^{-1}$. Show that the X_i are zero mean, uncorrelated, with variance $\sigma^2 = G^2 B \mathcal{N}_0$. Assuming the X_i are normal, and that $V_{100}^2 = 4.659$ mW, find the 95% confidence interval for $\mathcal{N}_0$; your answer should have units of Watts/Hz.

19. Let $X_1, X_2, \ldots$ be i.i.d. with known, finite mean m, unknown, finite variance σ^2, and finite, fourth central moment $\gamma^4 = \mathsf{E}[(X_i \hat{\Leftrightarrow} m)^4]$. Show that V_n^2 is a consistent estimator of σ^2. *Hint:* Apply Problem 7 with X_i replaced by $X_i \hat{\Leftrightarrow} m$.

20. Let $W_1, W_2, \ldots$ be i.i.d. $N(0, \sigma^2)$ with σ^2 unknown. Let $X_i = m + W_i$, where m is an unknown constant. Find the 95% confidence interval for σ^2 if $S_{100}^2 = 4.736$.

CHAPTER 13

Advanced Topics

Self similarity has been noted in the traffic of local-area networks [26], wide-area networks [32], and in World Wide Web traffic [12]. The purpose of this chapter is to introduce the notion of self similarity and related concepts so that students can be conversant with the kinds of stochastic processes being used to model network traffic.

Section 13.1 introduces the Hurst parameter and the notion of distributional self similarity for continuous-time processes. The concept of stationary increments is also presented. As an example of such processes, fractional Brownian motion is developed using the Wiener integral. In Section 13.2, we show that if one samples the increments of a continuous-time self-similar process with stationary increments, then the samples have a covariance function with a very specific formula. It is shown that this formula is equivalent to specifying the variance of the sample mean for all values of n. Also, the power spectral density is found up to a multiplicative constant. Section 13.3 introduces the concept of asymptotic second-order self similarity and shows that it is equivalent to specifying the limiting form of the variance of the sample mean. The main result here is a sufficient condition on the power spectral density that guarantees asymptotic second-order self similarity. Section 13.4 defines long-range dependence. It is shown that every long-range-dependent process is asymptotically second-order self similar. Section 13.5 introduces ARMA processes, and Section 13.6 extends this to fractional ARIMA processes. Fractional ARIMA process provide a large class of models that are asymptotically second-order self similar.

13.1. Self Similarity in Continuous Time

Loosely speaking, a continuous-time random process W_t is said to be **self similar** with **Hurst parameter** $H > 0$ if the process $W_{\lambda t}$ "looks like" the process $\lambda^H W_t$. If $\lambda > 1$, then time is speeded up for $W_{\lambda t}$ compared to the original process W_t. If $\lambda < 1$, then time is slowed down. The factor λ^H in $\lambda^H W_t$ either increases or decreases the magnitude (but not the time scale) compared with the original process. Thus, for a self-similar process, when time is speeded up, the apparent effect is the same as changing the magnitude of the original process, rather than its time scale.

The precise definition of "looks like" will be in the sense of finite-dimensional distributions. That is, for W_t to be self similar, we require that for every $\lambda > 0$, for every finite collection of times $t_1, \ldots, t_n$, all joint probabilities involving $W_{\lambda t_1}, \ldots, W_{\lambda t_n}$ are the same as those involving $\lambda^H W_{t_1}, \ldots, \lambda^H W_{t_n}$. The best example of a self-similar process is the Wiener process. This is easy to verify by comparing the joint characteristic functions of $W_{\lambda t_1}, \ldots, W_{\lambda t_n}$ and

$\lambda^H W_{t_1}, \ldots, \lambda^H W_{t_n}$ for the correct value of H.

Implications of Self Similarity

Let us focus first on a single time point t. For a self-similar process, we must have

$$\wp(W_{\lambda t} \le x) \;=\; \wp(\lambda^H W_t \le x).$$

Taking $t = 1$, results in

$$\wp(W_\lambda \le x) \;=\; \wp(\lambda^H W_1 \le x).$$

Since $\lambda > 0$ is a dummy variable, we can call it t instead. Thus,

$$\wp(W_t \le x) \;=\; \wp(t^H W_1 \le x), \quad t > 0.$$

Now rewrite this as

$$\wp(W_t \le x) \;=\; \wp(W_1 \le t^{-H} x), \quad t > 0,$$

or, in terms of cumulative distribution functions,

$$F_{W_t}(x) \;=\; F_{W_1}(t^{-H} x), \quad t > 0.$$

It can now be shown (Problem 1) that W_t converges in distribution to the zero random variable as $t \to 0$. Similarly, as $t \to \infty$, W_t converges in distribution to a discrete random variable taking the values 0 and $\pm\infty$.

We next look at expectations of self-similar processes. We can write

$$\mathsf{E}[W_{\lambda t}] \;=\; \mathsf{E}[\lambda^H W_t] \;=\; \lambda^H \mathsf{E}[W_t].$$

Setting $t = 1$ and replacing λ by t results in

$$\mathsf{E}[W_t] \;=\; t^H \mathsf{E}[W_1], \quad t > 0. \tag{13.1}$$

Hence, for a self-similar process, its mean function has the form of a constant times t^H for $t > 0$.

As another example, consider

$$\mathsf{E}[W_{\lambda t}^2] \;=\; \mathsf{E}\big[(\lambda^H W_t)^2\big] \;=\; \lambda^{2H} \mathsf{E}[W_t^2]. \tag{13.2}$$

Arguing as above, we find that

$$\mathsf{E}[W_t^2] \;=\; t^{2H} \mathsf{E}[W_1^2], \quad t > 0. \tag{13.3}$$

We can also take $t = 0$ in (13.2) to get

$$\mathsf{E}[W_0^2] \;=\; \lambda^{2H} \mathsf{E}[W_0^2].$$

Since the left-hand side does not depend on λ, $\mathsf{E}[W_0^2] = 0$, which implies $W_0 = 0$ a.s. Hence, (13.1) and (13.3) both continue to hold even when $t = 0$.*

*Using the formula $a^b = e^{b \ln a}$, $0^H = e^{H \ln 0} = e^{-\infty}$, since $H > 0$. Thus, $0^H = 0$.

***Example* 13.1.** Assuming that the Wiener process is self similar, show that the Hurst parameter must be $H = 1/2$.

Solution. Recall that for the Wiener process, we have $\mathsf{E}[W_t^2] = \sigma^2 t$. Thus, (13.2) implies that

$$\sigma^2(\lambda t) \;=\; \lambda^{2H}\sigma^2 t.$$

Hence, $H = 1/2$.

Stationary Increments

A process W_t is said to have **stationary increments** if for every increment $\tau > 0$, the **increment process**

$$Z_t \;:=\; W_t \Leftrightarrow W_{t-\tau}$$

is a stationary process in t. If W_t is self similar with Hurst parameter H, and has stationary increments, we say that W_t is ***H*-sssi**.

If Z_t is H-sssi, then $\mathsf{E}[Z_t]$ cannot depend on t; but by (13.1),

$$\mathsf{E}[Z_t] \;=\; \mathsf{E}[W_t \Leftrightarrow W_{t-\tau}] \;=\; \big[t^H \Leftrightarrow (t \Leftrightarrow \tau)^H\big]\mathsf{E}[W_1].$$

If $H \neq 1$, then we must have $\mathsf{E}[W_1] = 0$, which by (13.1), implies $\mathsf{E}[W_t] = 0$. As we will see later, the case $H = 1$ is not of interest if W_t has finite second moments, and so we will always take $\mathsf{E}[W_t] = 0$.

If Z_t is H-sssi, then the stationarity of the increments and (13.3) imply

$$\mathsf{E}[Z_t^2] \;=\; \mathsf{E}[Z_\tau^2] \;=\; \mathsf{E}[(W_\tau \Leftrightarrow W_0)^2] \;=\; \mathsf{E}[W_\tau^2] \;=\; \tau^{2H}\mathsf{E}[W_1^2].$$

Similarly, for $t > s$,

$$\mathsf{E}[(W_t \Leftrightarrow W_s)^2] = \mathsf{E}[(W_{t-s} \Leftrightarrow W_0)^2] = \mathsf{E}[W_{t-s}^2] = (t \Leftrightarrow s)^{2H}\mathsf{E}[W_1^2].$$

For $t < s$,

$$\mathsf{E}[(W_t \Leftrightarrow W_s)^2] \;=\; \mathsf{E}[(W_s \Leftrightarrow W_t)^2] \;=\; (s \Leftrightarrow t)^{2H}\mathsf{E}[W_1^2].$$

Thus, for arbitrary t and s,

$$\mathsf{E}[(W_t \Leftrightarrow W_s)^2] \;=\; |t \Leftrightarrow s|^{2H}\mathsf{E}[W_1^2].$$

Note in particular that

$$\mathsf{E}[W_t^2] \;=\; |t|^{2H}\mathsf{E}[W_1^2].$$

Now, we also have

$$\mathsf{E}[(W_t \Leftrightarrow W_s)^2] \;=\; \mathsf{E}[W_t^2] \Leftrightarrow 2\mathsf{E}[W_t W_s] + \mathsf{E}[W_s^2],$$

and it follows that

$$\mathsf{E}[W_t W_s] \;=\; \frac{\mathsf{E}[W_1^2]}{2}[|t|^{2H} \Leftrightarrow |t \Leftrightarrow s|^{2H} + |s|^{2H}]. \tag{13.4}$$

Fractional Brownian Motion

Let W_t denote the standard Wiener process on $\Leftrightarrow\infty < t < \infty$ as defined in Problem 23 in Chapter 8. The standard **fractional Brownian motion** is the process $B_H(t)$ defined by the Wiener integral

$$B_H(t) \;:=\; \int_{-\infty}^{\infty} g_{H,t}(\tau)\, dW_\tau,$$

where $g_{H,t}$ is defined below. Then $\mathsf{E}[B_H(t)] = 0$, and

$$\mathsf{E}[B_H(t)^2] \;=\; \int_{-\infty}^{\infty} g_{H,t}(\tau)^2\, d\tau.$$

To evaluate this expression as well as the correlation $\mathsf{E}[B_H(t)B_H(s)]$, we must now define $g_{H,t}(\tau)$. To this end, let

$$q_H(\theta) \;:=\; \begin{cases} \theta^{H-1/2}, & \theta > 0, \\ 0, & \theta \le 0, \end{cases}$$

and put

$$g_{H,t}(\tau) \;:=\; \frac{1}{C_H}\big[q_H(t \Leftrightarrow \tau) \Leftrightarrow q_H(\Leftrightarrow\tau)\big],$$

where

$$C_H^2 \;:=\; \int_0^{\infty} \big[(1+\theta)^{H-1/2} \Leftrightarrow \theta^{H-1/2}\big]^2\, d\theta + \frac{1}{2H}.$$

First note that since $g_{H,0}(\tau) = 0$, $B_H(0) = 0$. Next,

$$\begin{aligned} B_H(t) \Leftrightarrow B_H(s) &= \int_{-\infty}^{\infty} [g_{H,t}(\tau) \Leftrightarrow g_{H,s}(\tau)]\, dW_\tau \\ &= C_H^{-1} \int_{-\infty}^{\infty} [q_H(t \Leftrightarrow \tau) \Leftrightarrow q_H(s \Leftrightarrow \tau)]\, dW_\tau, \end{aligned}$$

and so

$$\mathsf{E}[|B_H(t) \Leftrightarrow B_H(s)|^2] \;=\; C_H^{-2} \int_{-\infty}^{\infty} |q_H(t \Leftrightarrow \tau) \Leftrightarrow q_H(s \Leftrightarrow \tau)|^2\, d\tau.$$

If we now assume $s < t$, then this integral is equal to the sum of

$$\int_{-\infty}^{s} [(t \Leftrightarrow \tau)^{H-1/2} \Leftrightarrow (s \Leftrightarrow \tau)^{H-1/2}]^2\, d\tau$$

and

$$\int_s^t (t \Leftrightarrow \tau)^{2H-1}\, d\tau \;=\; \int_0^{t-s} \theta^{2H-1}\, d\theta \;=\; \frac{(t \Leftrightarrow s)^{2H}}{2H}.$$

To evaluate the integral from $\Leftrightarrow\infty$ to s, let $\xi = s \Leftrightarrow \tau$ to get

$$\int_0^{\infty} [(t \Leftrightarrow s + \xi)^{H-1/2} \Leftrightarrow \xi^{H-1/2}]^2\, d\xi,$$

which is equal to

$$(t \Leftrightarrow s)^{2H-1} \int_0^\infty \left[\left(1 + \xi/(t \Leftrightarrow s)\right)^{H-1/2} \Leftrightarrow \left(\xi/(t \Leftrightarrow s)\right)^{H-1/2} \right]^2 d\xi.$$

Making the change of variable $\theta = \xi/(t \Leftrightarrow s)$ yields

$$(t \Leftrightarrow s)^{2H} \int_0^\infty \left[(1+\theta)^{H-1/2} \Leftrightarrow \theta^{H-1/2} \right]^2 d\theta.$$

It is now clear that

$$\mathsf{E}[|B_H(t) \Leftrightarrow B_H(s)|^2] \;=\; (t \Leftrightarrow s)^{2H}, \quad t > s.$$

Since interchanging the positions of t and s on the left-hand side has no effect, we can write for arbitrary t and s,

$$\mathsf{E}[|B_H(t) \Leftrightarrow B_H(s)|^2] \;=\; |t \Leftrightarrow s|^{2H}.$$

Taking $s = 0$ yields

$$\mathsf{E}[B_H(t)^2] \;=\; |t|^{2H}.$$

Furthermore, expanding $\mathsf{E}[|B_H(t) \Leftrightarrow B_H(s)|^2]$, we find that

$$|t \Leftrightarrow s|^{2H} \;=\; |t|^{2H} \Leftrightarrow 2\mathsf{E}[B_H(t)B_H(s)] + |s|^{2H},$$

or,

$$\mathsf{E}[B_H(t)B_H(s)] \;=\; \frac{|t|^{2H} \Leftrightarrow |t \Leftrightarrow s|^{2H} + |s|^{2H}}{2}. \tag{13.5}$$

Observe that $B_H(t)$ is a **Gaussian** process in the sense that if we select any sampling times, $t_1 < \cdots < t_n$, then the random vector $[B_H(t_1), \ldots, B_H(t_n)]'$ is Gaussian; this is a consequence of the fact that $B_H(t)$ is defined as a Wiener integral (Problem 14 in Chapter 11). Furthermore, the covariance matrix of the random vector is completely determined by (13.5). On the other hand, by (13.4), we see that *any* H-sssi process has the same covariance function (up to a scale factor). If that H-sssi process is Gaussian, then as far as the joint probabilities involving any finite number of sampling times, we may as well assume that the H-sssi process is fractional Brownian motion. In this sense, there is only one H-sssi process with finite second moments that is *Gaussian*: fractional Brownian motion.

13.2. Self Similarity in Discrete Time

Let W_t be an H-sssi process. By choosing an appropriate time scale for W_t, we can focus on the unit increment $\tau = 1$. Furthermore, the advent of digital signal processing suggests that we sample the increment process. This leads us to consider the discrete-time increment process

$$X_n \;:=\; W_n \Leftrightarrow W_{n-1}.$$

Since W_t is assumed to have zero mean, the covariance of X_n is easily found using (13.4). For $n > m$,

$$\mathsf{E}[X_n X_m] = \frac{\mathsf{E}[W_1^2]}{2}[(n \Leftrightarrow m + 1)^{2H} \Leftrightarrow 2(n \Leftrightarrow m)^{2H} + (n \Leftrightarrow m \Leftrightarrow 1)^{2H}].$$

Since this depends only on the time difference, the covariance function of X_n is

$$C(n) \;=\; \frac{\sigma^2}{2}[|n+1|^{2H} \Leftrightarrow 2|n|^{2H} + |n \Leftrightarrow 1|^{2H}], \tag{13.6}$$

where $\sigma^2 := \mathsf{E}[W_1^2]$.

The foregoing analysis assumed that X_n was obtained by sampling the increments of an H-sssi process. More generally, a discrete-time, wide-sense stationary (WSS) process is said to be **second-order self similar** if its covariance function has the form in (13.6). In this context it is not assumed that X_n is obtained from an underlying continuous-time process or that X_n has zero mean. A second-order self-similar process that is Gaussian is called **fractional Gaussian noise**, since one way of obtaining it is by sampling the increments of fractional Brownian motion.

Convergence Rates for the Mean-Square Law of Large Numbers

Suppose that X_n is a discrete-time, WSS process with mean $\mu := \mathsf{E}[X_n]$. It is shown later in Section 13.4 that if X_n is second-order self similar; i.e., if (13.6) holds, then $C(n) \to 0$. On account of Example 10.4, the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in mean square to μ. But how fast does the sample mean converge? We show that (13.6) holds if and only if

$$\mathsf{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i \Leftrightarrow \mu\right|^2\right] \;=\; \sigma^2 n^{2H-2} \;=\; \frac{\sigma^2}{n} n^{2H-1}. \tag{13.7}$$

In other words, X_n is second-order self similar if and only if (13.7) holds. To put (13.7) into perspective, first consider the case $H = 1/2$. Then (13.6) reduces to zero for $n \neq 0$. In other words, the X_n are uncorrelated. Also, when $H = 1/2$, the factor n^{2H-1} in (13.7) is not present. Thus, (13.7) reduces to the mean-square law of large numbers for uncorrelated random variables derived in Example 10.3. If $2H \Leftrightarrow 1 < 0$, or equivalently $H < 1/2$, then the convergence is faster than in the uncorrelated case. If $1/2 < H < 1$, then the convergence is slower than in the uncorrelated case. This has important consequences for determining confidence intervals, as shown in Problem 6.

To show the equivalence of (13.6) and (13.7), the first step is to define

$$Y_n \;:=\; \sum_{i=1}^{n} (X_i \Leftrightarrow \mu),$$

and observe that (13.7) is equivalent to $\mathsf{E}[Y_n^2] = \sigma^2 n^{2H}$.

The second step is to express $\mathsf{E}[Y_n^2]$ in terms of $C(n)$. Write

$$\begin{aligned}\mathsf{E}[Y_n^2] &= \mathsf{E}\left[\left(\sum_{i=1}^{n}(X_i \Leftrightarrow \mu)\right)\left(\sum_{k=1}^{n}(X_k \Leftrightarrow \mu)\right)\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{n}\mathsf{E}[(X_i \Leftrightarrow \mu)(X_k \Leftrightarrow \mu)] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{n} C(i \Leftrightarrow k). \end{aligned} \tag{13.8}$$

The above sum amounts to summing all the entries of the $n \times n$ matrix with $i\,k$ entry $C(i \Leftrightarrow k)$. This matrix is symmetric and is constant along each diagonal. Thus,

$$\mathsf{E}[Y_n^2] = nC(0) + 2\sum_{\nu=1}^{n-1} C(\nu)(n \Leftrightarrow \nu). \tag{13.9}$$

Now that we have a formula for $\mathsf{E}[Y_n^2]$ in terms of $C(n)$, we follow Likhanov [28, p. 195] and write

$$\mathsf{E}[Y_{n+1}^2] \Leftrightarrow \mathsf{E}[Y_n^2] = C(0) + 2\sum_{\nu=1}^{n} C(\nu),$$

and then

$$\big(\mathsf{E}[Y_{n+1}^2] \Leftrightarrow \mathsf{E}[Y_n^2]\big) \Leftrightarrow \big(\mathsf{E}[Y_n^2] \Leftrightarrow \mathsf{E}[Y_{n-1}^2]\big) = 2C(n). \tag{13.10}$$

Applying the formula $\mathsf{E}[Y_n^2] = \sigma^2 n^{2H}$ shows that for $n \geq 1$,

$$C(n) = \frac{\sigma^2}{2}[(n+1)^{2H} \Leftrightarrow 2n^{2H} + (n \Leftrightarrow 1)^{2H}].$$

Finally, it is a simple exercise (Problem 7) using induction on n to show that (13.6) implies $\mathsf{E}[Y_n^2] = \sigma^2 n^{2H}$ for $n \geq 1$.

Aggregation

Consider the partitioning of the sequence X_n into blocks of size m:

$$\underbrace{X_1, \ldots, X_m}_{\text{1st block}}\ \underbrace{X_{m+1}, \ldots, X_{2m}}_{\text{2nd block}} \cdots \underbrace{X_{(n-1)m+1}, \ldots, X_{nm}}_{n\text{th block}} \cdots.$$

The average of the first block is $\frac{1}{m}\sum_{k=1}^{m} X_k$. The average of the second block is $\frac{1}{m}\sum_{k=m+1}^{2m} X_k$. The average of the nth block is

$$X_n^{(m)} := \frac{1}{m}\sum_{k=(n-1)m+1}^{nm} X_k. \tag{13.11}$$

The superscript (m) indicates the block size, which is the number of terms used to compute the average. The subscript n indicates the block number. We call $\{X_n^{(m)}\}_{n=-\infty}^{\infty}$ the **aggregated process**. We now show that if X_n is second-order self similar, then the covariance function of $X_n^{(m)}$, denoted by $C^{(m)}(n)$, satisfies

$$C^{(m)}(n) \;=\; m^{2H-2}C(n). \tag{13.12}$$

In other words, if the original sequence is replaced by the sequence of averages of blocks of size m, then the new sequence has a covariance function that is the same as the original one except that the magnitude is scaled by m^{2H-2}.

The derivation of (13.12) is very similar to the derivation of (13.7). Put

$$\widetilde{X}_\nu^{(m)} \;:=\; \sum_{k=(\nu-1)m+1}^{\nu m} (X_k \Leftrightarrow \mu). \tag{13.13}$$

Since X_k is WSS, so is $\widetilde{X}_\nu^{(m)}$. Let its covariance function be denoted by $\widetilde{C}^{(m)}(\nu)$. Next define

$$\widetilde{Y}_n \;:=\; \sum_{\nu=1}^{n} \widetilde{X}_\nu^{(m)}.$$

Just as in (13.10),

$$2\widetilde{C}^{(m)}(n) \;=\; \big(\mathsf{E}[\widetilde{Y}_{n+1}^2] \Leftrightarrow \mathsf{E}[\widetilde{Y}_n^2]\big) \Leftrightarrow \big(\mathsf{E}[\widetilde{Y}_n^2] \Leftrightarrow \mathsf{E}[\widetilde{Y}_{n-1}^2]\big).$$

Now observe that

$$\begin{aligned}
\widetilde{Y}_n &= \sum_{\nu=1}^{n} \widetilde{X}_\nu^{(m)} \\
&= \sum_{\nu=1}^{n} \sum_{k=(\nu-1)m+1}^{\nu m} (X_k \Leftrightarrow \mu) \\
&= \sum_{\nu=1}^{nm} (X_\nu \Leftrightarrow \mu) \\
&= Y_{nm},
\end{aligned}$$

where this Y is the same as the one defined in the preceding subsection. Hence,

$$2\widetilde{C}^{(m)}(n) = \big(\mathsf{E}[Y_{(n+1)m}^2] \Leftrightarrow \mathsf{E}[Y_{nm}^2]\big) \Leftrightarrow \big(\mathsf{E}[Y_{nm}^2] \Leftrightarrow \mathsf{E}[Y_{(n-1)m}^2]\big). \tag{13.14}$$

Now we use the fact that since X_n is second-order self similar, $\mathsf{E}[Y_n^2] = \sigma^2 n^{2H}$. Thus,

$$\widetilde{C}^{(m)}(n) = \frac{\sigma^2}{2}\big[\big((n+1)m\big)^{2H} \Leftrightarrow 2(nm)^{2H} + \big((n \Leftrightarrow 1)m\big)^{2H}\big].$$

Since $C^{(m)}(n) = \widetilde{C}^{(m)}(n)/m^2$, (13.12) follows.

The Power Spectral Density

We show that the power spectral density of a second-order self-similar process is proportional to†

$$\sin^2(\pi f) \sum_{i=-\infty}^{\infty} \frac{1}{|i+f|^{2H+1}}. \tag{13.15}$$

The proof rests on the fact (derived below) that for any wide-sense stationary process,

$$\frac{C^{(m)}(n)}{m^{2H-2}} = \int_0^1 e^{j2\pi fn} \left(\sum_{k=0}^{m-1} \frac{S([f+k]/m)}{m^{2H+1}} \left[\frac{\sin(\pi f)}{\sin(\pi[f+k]/m)} \right]^2 \right) df$$

for $m = 1, 2, \ldots$. Now we also have

$$C(n) = \int_{-1/2}^{1/2} S(f) e^{j2\pi fn}\, df = \int_0^1 S(f) e^{j2\pi fn}\, df.$$

Hence, if $S(f)$ satisfies

$$\sum_{k=0}^{m-1} \frac{S([f+k]/m)}{m^{2H+1}} \left[\frac{\sin(\pi f)}{\sin(\pi[f+k]/m)} \right]^2 = S(f), \quad m = 1, 2, \ldots, \tag{13.16}$$

then (13.12) holds. In particular it holds for $n = 0$. Thus,

$$\frac{\mathsf{E}[Y_m^2]}{m^{2H}} = \frac{C^{(m)}(0)}{m^{2H-2}} = C(0) = \sigma^2, \quad m = 1, 2, \ldots.$$

As noted earlier, $\mathsf{E}[Y_n^2] = \sigma^2 n^{2H}$ is just (13.7), which is equivalent to (13.6). Now observe that if $S(f)$ is proportional to (13.15), then (13.16) holds.

The integral formula for $C^{(m)}(n)/m^{2H-2}$ is derived following Sinai [44, p. 66]. Write

$$\begin{aligned}
\frac{C^{(m)}(n)}{m^{2H-2}} &= \frac{1}{m^{2H-2}} \mathsf{E}[(X_{n+1}^{(m)} \Leftrightarrow \mu)(X_1^{(m)} \Leftrightarrow \mu)] \\
&= \frac{1}{m^{2H}} \sum_{i=nm+1}^{(n+1)m} \sum_{k=1}^{m} \mathsf{E}[(X_i \Leftrightarrow \mu)(X_k \Leftrightarrow \mu)] \\
&= \frac{1}{m^{2H}} \sum_{i=nm+1}^{(n+1)m} \sum_{k=1}^{m} C(i \Leftrightarrow k) \\
&= \frac{1}{m^{2H}} \sum_{i=nm+1}^{(n+1)m} \sum_{k=1}^{m} \int_{-1/2}^{1/2} S(f) e^{j2\pi f(i-k)}\, df \\
&= \frac{1}{m^{2H}} \int_{-1/2}^{1/2} S(f) \sum_{i=nm+1}^{(n+1)m} \sum_{k=1}^{m} e^{j2\pi f(i-k)}\, df.
\end{aligned}$$

†The constant of proportionality can be found using results from Section 13.3; see Problem 15.

Now write

$$\begin{aligned}\sum_{i=nm+1}^{(n+1)m}\sum_{k=1}^{m} e^{j2\pi f(i-k)} &= \sum_{\nu=1}^{m}\sum_{k=1}^{m} e^{j2\pi f(nm+\nu-k)} \\ &= e^{j2\pi nm}\sum_{\nu=1}^{m}\sum_{k=1}^{m} e^{j2\pi f(\nu-k)} \\ &= e^{j2\pi nm}\left|\sum_{k=1}^{m} e^{-j2\pi fk}\right|^2.\end{aligned}$$

Using the finite geometric series,

$$\sum_{k=1}^{m} e^{-j2\pi fk} = e^{-j2\pi f}\frac{1 \Leftrightarrow e^{-j2\pi fm}}{1 \Leftrightarrow e^{-j2\pi f}} = e^{-j\pi f(m+1)}\frac{\sin(m\pi f)}{\sin(\pi f)}.$$

Thus,

$$\frac{C^{(m)}(n)}{m^{2H-2}} = \frac{1}{m^{2H}}\int_{-1/2}^{1/2} S(f)e^{j2\pi fnm}\left[\frac{\sin(m\pi f)}{\sin(\pi f)}\right]^2 df.$$

Since the integrand has period one, we can shift the range of integration to $[0,1]$ and then make the change-of-variable $\theta = mf$. Thus,

$$\begin{aligned}\frac{C^{(m)}(n)}{m^{2H-2}} &= \frac{1}{m^{2H}}\int_{0}^{1} S(f)e^{j2\pi fnm}\left[\frac{\sin(m\pi f)}{\sin(\pi f)}\right]^2 df \\ &= \frac{1}{m^{2H+1}}\int_{0}^{m} S(\theta/m)e^{j2\pi\theta n}\left[\frac{\sin(\pi\theta)}{\sin(\pi\theta/m)}\right]^2 d\theta \\ &= \frac{1}{m^{2H+1}}\sum_{k=0}^{m-1}\int_{k}^{k+1} S(\theta/m)e^{j2\pi\theta n}\left[\frac{\sin(\pi\theta)}{\sin(\pi\theta/m)}\right]^2 d\theta.\end{aligned}$$

Now make the change-of-variable $f = \theta \Leftrightarrow k$ to get

$$\begin{aligned}\frac{C^{(m)}(n)}{m^{2H-2}} &= \frac{1}{m^{2H+1}}\sum_{k=0}^{m-1}\int_{0}^{1} S([f+k]/m)e^{j2\pi[f+k]n}\left[\frac{\sin(\pi[f+k])}{\sin(\pi[f+k]/m)}\right]^2 df \\ &= \frac{1}{m^{2H+1}}\sum_{k=0}^{m-1}\int_{0}^{1} S([f+k]/m)e^{j2\pi fn}\left[\frac{\sin(\pi f)}{\sin(\pi[f+k]/m)}\right]^2 df \\ &= \int_{0}^{1} e^{j2\pi fn}\left(\sum_{k=0}^{m-1}\frac{S([f+k]/m)}{m^{2H+1}}\left[\frac{\sin(\pi f)}{\sin(\pi[f+k]/m)}\right]^2\right) df.\end{aligned}$$

Notation

We have been using the term **correlation function** to refer to the quantity $\mathsf{E}[X_nX_m]$. This is the usual practice in engineering. However, engineers

studying network traffic follow the practice of statisticians and use the term **correlation function** to refer to

$$\frac{\mathsf{cov}(X_n, X_m)}{\sqrt{\mathsf{var}(X_n)\,\mathsf{var}(X_m)}}.$$

In other words, in networking, the term correlation function refers to our covariance function $C(n)$ divided by $C(0)$. We use the notation

$$\rho(n) \;:=\; \frac{C(n)}{C(0)}.$$

Now assume that X_n is second-order self similar. We have by (13.6) that $C(0) = \sigma^2$, and so

$$\rho(n) \;=\; \frac{1}{2}[|n+1|^{2H} \Leftrightarrow 2|n|^{2H} + |n \Leftrightarrow 1|^{2H}].$$

Let $\rho^{(m)}$ denote the correlation function of $X_n^{(m)}$. Then (13.12) tells us that

$$\rho^{(m)}(n) \;:=\; \frac{C^{(m)}(n)}{C^{(m)}(0)} \;=\; \frac{m^{2H-2}C(n)}{m^{2H-2}C(0)} \;=\; \rho(n).$$

13.3. Asymptotic Second-Order Self Similarity

We showed in the previous section that second-order self similarity (eq. (13.6)) is equivalent to (13.7), which specifies

$$\mathsf{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i \Leftrightarrow \mu\right|^2\right] \tag{13.17}$$

exactly for all n. While this is a very nice result, it applies only when the covariance function has exactly the form in (13.6). However, if we only need to know the behavior of (13.17) for large n, say

$$\lim_{n\to\infty} \frac{\mathsf{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i \Leftrightarrow \mu\right|^2\right]}{n^{2H-2}} \;=\; \sigma_\infty^2 \tag{13.18}$$

for some finite, positive σ_∞^2, then we can allow more freedom in the behavior of the covariance function. The key to obtaining such a result is suggested by (13.12), which says that for a second-order self similar process,

$$\frac{C^{(m)}(n)}{m^{2H-2}} \;=\; \frac{\sigma^2}{2}[|n+1|^{2H} \Leftrightarrow 2|n|^{2H} + |n \Leftrightarrow 1|^{2H}].$$

You are asked to show in Problems 9 and 10 that (13.18) holds if and only if

$$\lim_{m\to\infty} \frac{C^{(m)}(n)}{m^{2H-2}} \;=\; \frac{\sigma_\infty^2}{2}[|n+1|^{2H} \Leftrightarrow 2|n|^{2H} + |n \Leftrightarrow 1|^{2H}]. \tag{13.19}$$

A wide-sense stationary process that satisfies (13.19) is said to be **asymptotically** second-order self similar. In the literature, (13.19) is usually written in terms of the correlation function $\rho^{(m)}(n)$; see Problem 11.

If a wide-sense stationary process has a covariance function $C(n)$, how can we check if (13.18) or (13.19) holds? Below we answer this question in the frequency domain with a sufficient condition on the power spectral density. In Section 13.4, we answer this question in the time domain with a sufficient condition on $C(n)$ known as long-range dependence.

Let us look into the frequency domain. Suppose that $C(n)$ has a power spectral density $S(f)$ so that[‡]

$$C(n) \;=\; \int_{-1/2}^{1/2} S(f) e^{j2\pi fn}\, df.$$

The following result is proved later in this section.

Theorem. *If*

$$\lim_{f\to 0} \frac{S(f)}{|f|^{\alpha-1}} \;=\; s \tag{13.20}$$

for some finite, positive s, and if for every $0 < \delta < 1/2$, $S(f)$ is bounded on $[\delta, 1/2]$, then the process is asymptotically second-order self similar with $H = 1 \Leftrightarrow \alpha/2$ and

$$\sigma_\infty^2 \;=\; s \cdot \frac{4\cos(\alpha\pi/2),\,(\alpha)}{(2\pi)^\alpha\,(1 \Leftrightarrow \alpha)(2 \Leftrightarrow \alpha)}. \tag{13.21}$$

Notice that $0 < \alpha < 1$ implies $H = 1 \Leftrightarrow \alpha/2 \in (1/2, 1)$.

Below we give a specific power spectral density that satisfies the above conditions.

***Example* 13.2** (Hosking [23, Theorem 1(c)]). Fix $0 < d < 1/2$, and let[§]

$$S(f) \;=\; |1 \Leftrightarrow e^{-j2\pi f}|^{-2d}.$$

Since

$$1 \Leftrightarrow e^{-j2\pi f} \;=\; 2je^{-j\pi f}\,\frac{e^{j\pi f} \Leftrightarrow e^{-j\pi f}}{2j},$$

we can write

$$S(f) \;=\; [4\sin^2(\pi f)]^{-d}.$$

[‡] The power spectral density of a discrete-time process is periodic with period 1, and is real, even, and nonnegative. It is integrable since

$$\int_{-1/2}^{1/2} S(f)\,df \;=\; C(0) \;<\; \infty.$$

[§] As shown in Section 13.6, a process with this power spectral density is an ARIMA$(0, d, 0)$ process. The covariance function that corresponds to $S(f)$ is derived in Problem 12.

Since

$$\lim_{f\to 0} \frac{[4\sin^2(\pi f)]^{-d}}{[4(\pi f)^2]^{-d}} \;=\; 1,$$

if we put $\alpha = 1 \Leftrightarrow 2d$, then

$$\lim_{f\to 0} \frac{S(f)}{|f|^{\alpha-1}} \;=\; (2\pi)^{-2d}.$$

Notice that to keep $0 < \alpha < 1$, we needed $0 < d < 1/2$.

The power spectral density $S(f)$ in the above example factors into

$$S(f) \;=\; [1 \Leftrightarrow e^{-j2\pi f}]^{-d}[1 \Leftrightarrow e^{j2\pi f}]^{-d}.$$

More generally, let $S(f)$ be any power spectral density satisfying (13.20), boundedness away from the origin, and having a factorization of the form $S(f) = G(f)G(f)^*$. Then pass any wide-sense stationary, uncorrelated sequence though the discrete-time filter $G(f)$. By the discrete-time analog of (6.6), the output power spectral density is proportional to $S(f)$, and therefore asymptotically second-order self similar.

***Example* 13.3** (Hosking [23, Theorem 1(a)]). Find the impulse response of the filter $G(f) = [1 \Leftrightarrow e^{-j2\pi f}]^{-d}$.

Solution. Observe that $G(f)$ can be obtained by evaluating the z transform $(1 \Leftrightarrow z^{-1})^{-d}$ on the unit circle, $z = e^{j2\pi f}$. Hence, the desired impulse response can be found by inspection of the series for $(1 \Leftrightarrow z^{-1})^{-d}$. To this end, it is easy to show that the Taylor series for $(1+z)^d$ is¶

$$(1+z)^d \;=\; 1 + \sum_{n=1}^{\infty} \frac{d(d \Leftrightarrow 1)\cdots(d \Leftrightarrow [n \Leftrightarrow 1])}{n!} z^n.$$

Hence,

$$\begin{aligned}(1 \Leftrightarrow z^{-1})^{-d} &= 1 + \sum_{n=1}^{\infty} \frac{(\Leftrightarrow d)(\Leftrightarrow d \Leftrightarrow 1)\cdots(\Leftrightarrow d \Leftrightarrow [n \Leftrightarrow 1])}{n!} (\Leftrightarrow z^{-1})^n \\ &= 1 + \sum_{n=1}^{\infty} \frac{d(d+1)\cdots(d+[n \Leftrightarrow 1])}{n!} z^{-n}.\end{aligned}$$

Note that the impulse response is causal.

¶Notice that if $d \geq 0$ is an integer, the product

$$d(d-1)\cdots(d-[n-1])$$

contains zero as a factor for $n \geq d+1$; in this case, the sum contains only $d+1$ terms and converges for all complex z. In fact, the formula reduces to the binomial theorem.

Once we have a process whose power spectral density satisfies (13.20) and boundedness away from the origin, it remains so after further filtering by *stable* linear time-invariant systems. For if $\sum_n |h_n| < \infty$, then

$$H(f) \;=\; \sum_{n=-\infty}^{\infty} h_n e^{-j2\pi fn}$$

is an absolutely convergent series and therefore continuous. If $S(f)$ satisfies (13.20), then

$$\lim_{f\to 0} \frac{|H(f)|^2 S(f)}{|f|^{\alpha-1}} \;=\; |H(0)|^2 s.$$

A wide class of stable filters is provided by autoregressive moving average (ARMA) systems discussed in Section 13.5.

Proof of Theorem. We now establish that (13.20) and boundedness of $S(f)$ away from the origin imply asymptotic second-order self similarity. Since (13.19) and (13.18) are equivalent, it suffices to establish (13.18). As noted in the Hint in Problem 10, (13.18) is equivalent to $\mathsf{E}[Y_n^2]/n^{2H} \to \sigma_\infty^2$. From (13.8),

$$\begin{aligned}
\mathsf{E}[Y_n^2] &= \sum_{i=1}^{n}\sum_{k=1}^{n} C(i \ominus k) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{n} \int_{-1/2}^{1/2} S(f) e^{j2\pi f(i-k)}\, df \\
&= \int_{-1/2}^{1/2} S(f) \left|\sum_{k=1}^{n} e^{-j2\pi fk}\right|^2 df.
\end{aligned}$$

Using the finite geometric series,

$$\sum_{k=1}^{n} e^{-j2\pi fk} \;=\; e^{-j2\pi f}\,\frac{1 \ominus e^{-j2\pi fn}}{1 \ominus e^{-j2\pi f}} \;=\; e^{-j\pi f(n+1)}\,\frac{\sin(n\pi f)}{\sin(\pi f)}.$$

Thus,

$$\begin{aligned}
\mathsf{E}[Y_n^2] &= \int_{-1/2}^{1/2} S(f) \left[\frac{\sin(n\pi f)}{\sin(\pi f)}\right]^2 df \\
&= n^2 \int_{-1/2}^{1/2} S(f) \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 \left[\frac{\pi f}{\sin(\pi f)}\right]^2 df.
\end{aligned}$$

We now show that

$$\lim_{n\to\infty} \frac{\mathsf{E}[Y_n^2]}{n^{2-\alpha}} \;=\; s \cdot \frac{4\cos(\alpha\pi/2),\ (\alpha)}{(2\pi)^\alpha\,(1 \ominus \alpha)\,(2 \ominus \alpha)}.$$

The first step is to put

$$K_n(\alpha) \;:=\; n^{\alpha} \int_{-1/2}^{1/2} \frac{1}{|f|^{1-\alpha}} \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 \left[\frac{\pi f}{\sin(\pi f)}\right]^2 df,$$

and show that

$$\frac{\mathsf{E}[Y_n^2]}{n^{2-\alpha}} \Leftrightarrow s \cdot K_n(\alpha) \;\to\; 0. \tag{13.22}$$

The second step, which is left to Problem 13, is to show that $K_n(\alpha) \to K(\alpha)$, where

$$K(\alpha) \;:=\; \pi^{-\alpha} \int_{-\infty}^{\infty} \frac{1}{|\theta|^{1-\alpha}} \Big(\frac{\sin\theta}{\theta}\Big)^2 d\theta.$$

The third step, which is left to Problem 14, is to show that

$$K(\alpha) \;=\; \frac{4\cos(\alpha\pi/2),\ (\alpha)}{(2\pi)^{\alpha}(1 \Leftrightarrow \alpha)(2 \Leftrightarrow \alpha)}.$$

To begin, observe that

$$\frac{\mathsf{E}[Y_n^2]}{n^{2-\alpha}} \Leftrightarrow s \cdot K_n(\alpha)$$

is equal to

$$n^{\alpha} \int_{-1/2}^{1/2} \left[S(f) \Leftrightarrow \frac{s}{|f|^{1-\alpha}}\right] \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 \left[\frac{\pi f}{\sin(\pi f)}\right]^2 df. \tag{13.23}$$

Let $\varepsilon > 0$ be given, and let $0 < \delta < 1/2$ be so small that

$$\Big|\frac{S(f)}{|f|^{\alpha-1}} \Leftrightarrow s\Big| \;<\; \varepsilon,$$

or

$$\Big|S(f) \Leftrightarrow \frac{s}{|f|^{1-\alpha}}\Big| \;<\; \frac{\varepsilon}{|f|^{1-\alpha}}.$$

In analyzing (13.23), it is first convenient to focus on the range of integration $[0, \delta]$. Then the absolute value of

$$n^{\alpha} \int_0^{\delta} \left[S(f) \Leftrightarrow \frac{s}{|f|^{1-\alpha}}\right] \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 \left[\frac{\pi f}{\sin(\pi f)}\right]^2 df$$

is upper bounded by

$$\varepsilon n^{\alpha} \Big(\frac{\pi}{2}\Big)^2 \int_0^{\delta} \frac{1}{f^{1-\alpha}} \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 df,$$

where we have used the fact that for $|f| \le 1/2$,

$$1 \;\ge\; \frac{\sin(\pi f)}{\pi f} \;\ge\; \frac{2}{\pi}.$$

Now make the change-of-variable $\theta = n\pi f$ in the above integral to get the expression

$$\varepsilon n^{\alpha}\Big(\frac{\pi}{2}\Big)^2 \frac{\pi^{-\alpha}}{n^{\alpha}} \int_0^{n\pi\delta} \frac{1}{\theta^{1-\alpha}} \left[\frac{\sin\theta}{\theta}\right]^2 d\theta,$$

or

$$\varepsilon\Big(\frac{\pi}{2}\Big)^2 \pi^{-\alpha} \int_0^{n\pi\delta} \frac{1}{\theta^{1-\alpha}} \left[\frac{\sin\theta}{\theta}\right]^2 d\theta \;\le\; \varepsilon\Big(\frac{\pi}{2}\Big)^2 \frac{K(\alpha)}{2}.$$

We return to (13.23) and focus now on the range of integration $[\delta, 1/2]$. The absolute value of

$$n^{\alpha} \int_{\delta}^{1/2} \left[S(f) \Leftrightarrow \frac{s}{|f|^{1-\alpha}}\right] \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 \left[\frac{\pi f}{\sin(\pi f)}\right]^2 df$$

is upper bounded by

$$n^{\alpha}\Big(\frac{\pi}{2}\Big)^2 B_{\delta} \int_{\delta}^{1/2} \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 df, \tag{13.24}$$

where

$$B_{\delta} \;:=\; \sup_{|f|\in[\delta,1/2]} \left| S(f) \Leftrightarrow \frac{s}{|f|^{1-\alpha}} \right|.$$

In (13.24) make the change-of-variable $\theta = n\pi f$ to get

$$\begin{aligned} \int_{\delta}^{1/2} \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 df &= \int_{n\pi\delta}^{n\pi/2} \left[\frac{\sin\theta}{\theta}\right]^2 \frac{d\theta}{n\pi} \\ &\le \int_0^{\infty} \left[\frac{\sin\theta}{\theta}\right]^2 \frac{d\theta}{n\pi}. \end{aligned}$$

Thus, (13.24) is upper bounded by

$$n^{\alpha-1}\Big(\frac{\pi}{4}\Big) B_{\delta} \int_0^{\infty} \left[\frac{\sin\theta}{\theta}\right]^2 d\theta.$$

Since this integral is finite, and since $\alpha < 1$, this bound goes to zero as $n \to \infty$. Thus, (13.22) is established.

13.4. Long-Range Dependence

Loosely speaking, a wide-sense stationary process is said to be **long-range dependent** (**LRD**) if its covariance function $C(n)$ decays slowly as $n \to \infty$. The precise definition of slow decay is the requirement that for some $0 < \alpha < 1$,

$$\lim_{n\to\infty} \frac{C(n)}{n^{-\alpha}} \;=\; c, \tag{13.25}$$

for some finite, positive constant c. In other words, for large n, $C(n)$ looks like c/n^{α}.

In this section, we prove two important results. The first result is that a second-order self-similar process is long-range dependent. The second result is that long-range dependence implies asymptotic second-order self similarity.

To prove that second-order self similarity implies long-range dependence, we proceed as follows. Write (13.6) for $n \geq 1$ as

$$C(n) \;=\; \frac{\sigma^2}{2} n^{2H} [(1+1/n)^{2H} \Leftrightarrow 2 + (1 \Leftrightarrow 1/n)^{2H}] \;=\; \frac{\sigma^2}{2} n^{2H} q(1/n),$$

where

$$q(t) \;:=\; (1+t)^{2H} \Leftrightarrow 2 + (1 \Leftrightarrow t)^{2H}.$$

For n large, $1/n$ is small. This suggests that we examine the Taylor expansion of $q(t)$ for t near zero. Since $q(0) = q'(0) = 0$, we expand to second order to get

$$q(t) \;\approx\; \frac{q''(0)}{2} t^2 \;=\; 2H(2H \Leftrightarrow 1) t^2.$$

So, for large n,

$$C(n) \;=\; \frac{\sigma^2}{2} n^{2H} q(1/n) \;\approx\; \sigma^2 H(2H \Leftrightarrow 1) n^{2H-2}. \tag{13.26}$$

It appears that $\alpha = 2 \Leftrightarrow 2H$ and $c = \sigma^2 H(2H \Leftrightarrow 1)$. Note that $0 < \alpha < 1$ corresponds to $1/2 < H < 1$. Also, $H > 1/2$ corresponds to $\sigma^2 H(2H \Leftrightarrow 1) > 0$. To prove that these values of α and c work, write

$$\lim_{n\to\infty} \frac{C(n)}{n^{2H-2}} \;=\; \frac{\sigma^2}{2} \lim_{n\to\infty} \frac{q(1/n)}{n^{-2}} \;=\; \frac{\sigma^2}{2} \lim_{t \downarrow 0} \frac{q(t)}{t^2},$$

and apply l'Hôpital's rule twice to obtain

$$\lim_{n\to\infty} \frac{C(n)}{n^{2H-2}} \;=\; \sigma^2 H(2H \Leftrightarrow 1). \tag{13.27}$$

This formula implies the following two facts. First, if $H > 1$, then $C(n) \to \infty$ as $n \to \infty$. This contradicts the fact that covariance functions are bounded (recall that $|C(n)| \leq C(0)$ by the Cauchy–Schwarz inequality; cf. Section 6.2). Thus, a second-order self-similar process cannot have $H > 1$. Second, if $H = 1$, then $C(n) \to \sigma^2$. In other words, the covariance does not decay to zero as n increases. Since this situation does not arise in applications, we do not consider the case $H = 1$.

We now show that long-range dependence (13.25) implies[‖] asymptotic second-order self similarity with $H = 1 \Leftrightarrow \alpha/2$ and $\sigma_\infty^2 = 2c/[(1 \Leftrightarrow \alpha)(2 \Leftrightarrow \alpha)]$. From

[‖]Actually, the weaker condition,

$$\lim_{n\to\infty} \frac{C(n)}{\ell(n) n^{-\alpha}} \;=\; c,$$

where ℓ is a **slowly varying function**, is enough to imply asymptotic second-order self similarity [48]. The derivation we present results from taking the proof in [48, Appendix A] and setting $\ell(n) \equiv 1$ so that no theory of slowly varying functions is required.

(13.9),

$$\frac{\mathsf{E}[Y_n^2]}{n^{2-\alpha}} \;=\; \frac{C(0)}{n^{1-\alpha}} + 2\frac{\sum_{\nu=1}^{n-1} C(\nu)}{n^{1-\alpha}} \Leftrightarrow 2\frac{\sum_{\nu=1}^{n-1} \nu\, C(\nu)}{n^{2-\alpha}}.$$

We claim that if (13.25) holds, then

$$\lim_{n\to\infty} \frac{\sum_{\nu=1}^{n-1} C(\nu)}{n^{1-\alpha}} \;=\; \frac{c}{1\Leftrightarrow\alpha}, \tag{13.28}$$

and

$$\lim_{n\to\infty} \frac{\sum_{\nu=1}^{n-1} \nu\, C(\nu)}{n^{2-\alpha}} \;=\; \frac{c}{2\Leftrightarrow\alpha}. \tag{13.29}$$

Since $n^{1-\alpha}\to\infty$, it follows that

$$\lim_{n\to\infty} \frac{\mathsf{E}[Y_n^2]}{n^{2-\alpha}} \;=\; \frac{2c}{1\Leftrightarrow\alpha} \Leftrightarrow \frac{2c}{2\Leftrightarrow\alpha} \;=\; \frac{2c}{(1\Leftrightarrow\alpha)(2\Leftrightarrow\alpha)}.$$

Since $n^{1-\alpha}\to\infty$, to prove (13.28), it is enough to show that for some k,

$$\lim_{n\to\infty} \frac{\sum_{\nu=k}^{n-1} C(\nu)}{n^{1-\alpha}} \;=\; \frac{c}{1\Leftrightarrow\alpha}.$$

Fix any $0<\varepsilon<c$. By (13.25), there is a k such that for all $\nu\ge k$,

$$\left|\frac{C(\nu)}{\nu^{-\alpha}} \Leftrightarrow c\right| \;<\; \varepsilon.$$

Then

$$(c\Leftrightarrow\varepsilon)\sum_{\nu=k}^{n-1}\nu^{-\alpha} \;\le\; \sum_{\nu=k}^{n-1} C(\nu) \;\le\; (c+\varepsilon)\sum_{\nu=k}^{n-1}\nu^{-\alpha}.$$

Hence, we only need to prove that

$$\lim_{n\to\infty} \frac{\sum_{\nu=k}^{n-1} \nu^{-\alpha}}{n^{1-\alpha}} \;=\; \frac{1}{1\Leftrightarrow\alpha}.$$

This is done in Problem 17 by exploiting the inequality

$$\sum_{\nu=k}^{n-1}(\nu+1)^{-\alpha} \;\le\; \int_k^n t^{-\alpha}\,dt \;\le\; \sum_{\nu=k}^{n-1}\nu^{-\alpha}. \tag{13.30}$$

Note that

$$I_n \;:=\; \int_k^n t^{-\alpha}\,dt \;=\; \frac{n^{1-\alpha}\Leftrightarrow k^{1-\alpha}}{1\Leftrightarrow\alpha} \;\to\; \infty \quad \text{as } n\to\infty.$$

A similar approach is used in Problem 18 to derive (13.29).

13.5. ARMA Processes

We say that X_n is an **autoregressive moving average** (ARMA) process if X_n satisfies the equation

$$X_n + a_1 X_{n-1} + \cdots + a_p X_{n-p} \;=\; Z_n + b_1 Z_{n-1} + \cdots + b_q Z_{n-q}, \qquad (13.31)$$

where Z_n is an uncorrelated sequence of zero-mean random variables with common variance $\sigma^2 = \mathsf{E}[Z_n]$. In this case, we say that X_n is ARMA(p,q). If $a_1 = \cdots = a_p = 0$, then

$$X_n \;=\; Z_n + b_1 Z_{n-1} + \cdots + b_q Z_{n-q},$$

and we say that X_n is a **moving average** process, denoted by MA(q). If instead $b_1 = \cdots = b_q = 0$, then

$$X_n \;=\; \Leftrightarrow(a_1 X_{n-1} + \cdots + a_p X_{n-p}),$$

and we say that X_n is an **autoregressive** process, denoted by AR(p).

To gain some insight into (13.31), rewrite it using convolution sums as

$$\sum_{k=-\infty}^{\infty} a_k X_{n-k} \;=\; \sum_{k=-\infty}^{\infty} b_k Z_{n-k}, \qquad (13.32)$$

where

$$a_0 \;:=1, \quad a_k \;:=\; 0, \quad k<0 \text{ and } k>p,$$

and

$$b_0 \;:=1, \quad b_k \;:=\; 0, \quad k<0 \text{ and } k>q.$$

Taking z transforms of (13.32) yields

$$A(z)X(z) \;=\; B(z)Z(z),$$

or

$$X(z) \;=\; \frac{B(z)}{A(z)} Z(z),$$

where

$$A(z) \;:=\; 1 + a_1 z^{-1} + \cdots + a_p z^{-p} \quad \text{and} \quad B(z) \;:=\; 1 + b_1 z^{-1} + \cdots + b_q z^{-q}.$$

This suggests that if h_n has z transform $H(z) := B(z)/A(z)$, and if

$$X_n \;:=\; \sum_k h_k Z_{n-k} \;=\; \sum_k h_{n-k} Z_k, \qquad (13.33)$$

then (13.32) holds. This is indeed the case, as can be seen by writing

$$\begin{aligned}
\sum_i a_i X_{n-i} \;&=\; \sum_i a_i \sum_k h_{n-i-k} Z_k \\
&=\; \sum_k \left(\sum_i a_i h_{n-k-i} \right) Z_k \\
&=\; \sum_k b_{n-k} Z_k,
\end{aligned}$$

since $A(z)H(z) = B(z)$.

The "catch" in the preceding argument is to make sure that the infinite sums in (13.33) are well defined. If h_n is causal ($h_n = 0$ for $n < 0$), and if h_n is stable ($\sum_n |h_n| < \infty$), then (13.33) holds in L^2, L^1, and almost surely (recall Example 10.7 and Problems 17 and 32 in Chapter 11). Hence, it remains to prove the key result of this section, that if $A(z)$ has all roots strictly inside the unit circle, then h_n is causal and stable.

To begin the proof, observe that since $A(z)$ has all its roots inside the unit circle, the polynomial $\alpha(z) := A(1/z)$ has all its roots strictly outside the unit circle. Hence, for small enough $\delta > 0$, $1/\alpha(z)$ has the power series expansion

$$\frac{1}{\alpha(z)} \;=\; \sum_{n=0}^{\infty} \alpha_n z^n, \quad |z| < 1+\delta,$$

for unique coefficients α_n. In particular, this series converges for $z = 1 + \delta/2$. Since the terms of a convergent series go to zero, we must have $\alpha_n(1+\delta/2)^n \to 0$. Since a convergent sequence is bounded, there is some finite M for which $|\alpha_n(1+\delta/2)^n| \le M$, or $|\alpha_n| \le M(1+\delta/2)^{-n}$, which is summable by the geometric series. Thus, $\sum_n |\alpha_n| < \infty$. Now write

$$H(z) \;=\; \frac{B(z)}{A(z)} \;=\; \frac{B(z)}{\alpha(1/z)} \;=\; B(z)\frac{1}{\alpha(1/z)},$$

or

$$H(z) \;=\; B(z)\sum_{n=0}^{\infty} \alpha_n z^{-n} \;=\; \sum_{n=-\infty}^{\infty} h_n z^{-n},$$

where h_n is given by the convolution

$$h_n \;=\; \sum_{k=-\infty}^{\infty} \alpha_k b_{n-k}.$$

Since α_n and b_n are causal, so is their convolution h_n. Furthermore, for $n \ge 0$,

$$h_n \;=\; \sum_{k=\max(0,n-q)}^{n} \alpha_k b_{n-k}. \tag{13.34}$$

In Problem 19, you are asked to show that $\sum_n |h_n| < \infty$. In Problem 20, you are asked to show that (13.33) is the *unique* solution of (13.32).

13.6. ARIMA Processes

Before defining ARIMA processes, we introduce the **differencing filter**, whose z transform is $1 \Leftrightarrow z^{-1}$. If the input to this filter is X_n, then the output is $X_n \Leftrightarrow X_{n-1}$.

A process X_n is said to be an **autoregressive integrated moving average** (ARIMA) process if instead of $A(z)X(z) = B(z)Z(z)$, we have

$$A(z)(1 \Leftrightarrow z^{-1})^d X(z) \;=\; B(z)Z(z), \tag{13.35}$$

where $A(z)$ and $B(z)$ are defined as in the previous section. In this case, we say that X_n is an ARIMA(p,d,q) process. If we let $\widetilde{A}(z) = A(z)(1 \Leftrightarrow z^{-1})^d$, it would seem that ARIMA(p,d,q) is just a fancy name for ARMA$(p+d,q)$. While this is true when d is a nonnegative integer, there are two problems. First, recall that the results of the previous section assume $A(z)$ has all roots strictly inside the unit circle, while $\widetilde{A}(z)$ has a root at $z=1$ repeated d times. The second problem is that we will be focusing on fractional values of d, in which case $\widetilde{A}(1/z)$ is no longer a polynomial, but an infinite power series in z.

Let us rewrite (13.35) as

$$X(z) \;=\; (1 \Leftrightarrow z^{-1})^{-d} \frac{B(z)}{A(z)} Z(z) \;=\; H(z)G_d(z)Z(z),$$

where $H(z) := B(z)/A(z)$ as in the previous section, and

$$G_d(z) \;:=\; (1 \Leftrightarrow z^{-1})^{-d}.$$

From the calculations following Example 13.2,**

$$G_d(z) \;=\; \sum_{n=0}^{\infty} g_n z^{-n},$$

where $g_0 = 1$, and for $n \geq 1$,

$$g_n \;=\; \frac{d(d+1)\cdots(d+[n \Leftrightarrow 1])}{n!}.$$

The plan then is to set

$$Y_n \;:=\; \sum_{k=0}^{\infty} g_k Z_{n-k}$$

and then††

$$X_n \;:=\; \sum_{k=0}^{\infty} h_k Y_{n-k}.$$

Note that the power spectral density of Y is‡‡

$$S_Y(f) \;=\; |G_d(e^{j2\pi f})|^2 \sigma^2 \;=\; |1 \Leftrightarrow e^{-j2\pi f}|^{-2d}\sigma^2 \;=\; [4\sin^2(\pi f)]^{-d}\sigma^2,$$

Since $1 - z^{-1}$ is a differencing filter, $(1-z^{-1})^{-1}$ is a summing or **integrating filter. For noninteger values of d, $(1-z^{-1})^{-d}$ is called a *fractional* integrating filter. The corresponding process is sometimes called a **fractional ARIMA process** (FARIMA).

††Recall that h_n is given by (13.34).

‡‡We are appealing to the discrete-time version of (6.6).

using the result of Example 13.2. If $p = q = 0$, then $A(z) = B(z) = H(z) = 1$, $X_n = Y_n$, and we see that the process of Example 13.2 is ARIMA$(0, d, 0)$.

Now, the problem with the above plan is that we have to make sure that Y_n is well defined. To analyze the situation, we need to know how fast the g_n decay. To this end, observe that

$$\begin{aligned} \Gamma(d+n) &= (d+[n \Leftrightarrow 1])\,\Gamma(d+[n \Leftrightarrow 1]) \\ &\vdots \\ &= (d+[n \Leftrightarrow 1]) \cdots (d+1)\,\Gamma(d+1). \end{aligned}$$

Hence,

$$g_n = \frac{d \cdot \Gamma(d+n)}{\Gamma(d+1)\,\Gamma(n+1)}.$$

Now apply Stirling's formula,*

$$\Gamma(x) \sim \sqrt{2\pi}\, x^{x-1/2} e^{-x},$$

to the gamma functions that involve n. This yields

$$g_n \sim \frac{d e^{1-d}}{\Gamma(d+1)} \Big(1 + \frac{d \Leftrightarrow 1}{n+1}\Big)^{n+1/2} (n+d)^{d-1}.$$

Since

$$\Big(1 + \frac{d \Leftrightarrow 1}{n+1}\Big)^{n+1/2} = \Big(1 + \frac{d \Leftrightarrow 1}{n+1}\Big)^{n+1} \Big(1 + \frac{d \Leftrightarrow 1}{n+1}\Big)^{-1/2} \to e^{d-1},$$

and since $(n+d)^{d-1} \sim n^{d-1}$, we see that

$$g_n \sim \frac{d}{\Gamma(d+1)} n^{d-1},$$

as in Hosking [23, Theorem 1(a)]. For $0 < d < 1/2$, $\Leftrightarrow 1 < d \Leftrightarrow 1 < \Leftrightarrow 1/2$, and we see that the g_n are not absolutely summable. However, since $\Leftrightarrow 2 < 2d \Leftrightarrow 2 < \Leftrightarrow 1$, they are square summable. Hence, Y_n is well defined as a limit in mean square by Problem 26 in Chapter 11. The sum defining X_n is well defined in L^2, L^1, and almost surely by Example 10.7 and Problems 17 and 32 in Chapter 11. Since X_n is the result of filtering the long-range dependent process Y_n with the stable impulse response h_n, X_n is still long-range dependent as pointed out in Section 13.4.

*We derived Stirling's formula for $\Gamma(n) = (n-1)!$ in Example 4.14. A proof for noninteger x can be found in [6, pp. 300–301].

13.7. Problems

Problems §13.1: Self Similarity in Continuous Time

1. Show that for a self-similar process, W_t converges in distribution to the zero random variable as $t \to 0$. Next, identify $\lim_{t\to\infty} t^H W_1(\omega)$ as a function of ω, and find the probability mass function of the limit in terms of $F_{W_1}(x)$.

2. Use joint characteristic functions to show that the Wiener process is self similar with Hurst parameter $H = 1/2$.

3. Use joint characteristic functions to show that the Wiener process has stationary increments.

4. Show that for $H = 1/2$,

$$B_H(t) \Leftrightarrow B_H(s) \;=\; \int_s^t dW_\tau \;=\; W_t \Leftrightarrow W_s.$$

Taking $t > s = 0$ shows that $B_H(t) = W_t$, while taking $s < t = 0$ shows that $B_H(s) = W_s$. Thus, $B_{1/2}(t) \;=\; W_t$ for all t.

5. Show that for $0 < H < 1$,

$$\int_0^\infty \left[(1+\theta)^{H-1/2} \Leftrightarrow \theta^{H-1/2}\right]^2 d\theta \;<\; \infty.$$

Problems §13.2: Self Similarity in Discrete Time

6. Let X_n be a second-order self-similar process with mean $\mu = \mathsf{E}[X_n]$, variance $\sigma^2 = \mathsf{E}[(X_n \Leftrightarrow \mu)^2]$, and Hurst parameter H. Then the sample mean

$$M_n \;:=\; \frac{1}{n}\sum_{i=1}^n X_i$$

has expectation μ and, by (13.7), variance σ^2/n^{2-2H}. If X_n is a Gaussian sequence,

$$\frac{M_n \Leftrightarrow \mu}{\sigma/n^{1-H}} \;\sim\; N(0,1),$$

and so given a confidence level $1 \Leftrightarrow \alpha$, we can choose y (e.g., by Table 12.1) such that

$$\wp\Big(\Big|\frac{M_n \Leftrightarrow \mu}{\sigma/n^{1-H}}\Big| \le y\Big) \;=\; 1 \Leftrightarrow \alpha.$$

For $1/2 < H < 1$, show that the width of the corresponding confidence interval is wider by a factor of $n^{H-1/2}$ than the confidence interval obtained if the X_n had been independent as in Section 12.2.

7. Use (13.10) and induction on n to show that (13.6) implies $\mathsf{E}[Y_n^2] = \sigma^2 n^{2H}$.

8. Suppose that X_k is wide-sense stationary.

   (a) Show that the process $\widetilde{X}_\nu^{(m)}$ defined in (13.13) is also wide-sense stationary.

   (b) If X_k is second-order self similar, prove (13.12) for the case $n = 0$.

Problems §13.3: Asymptotic Second-Order Self Similarity

9. Show that asymptotic second-order self similarity (13.19) implies (13.18). *Hint:* Observe that $C^{(n)}(0) = \mathsf{E}[(X_1^{(n)} \Leftrightarrow \mu)^2]$.

10. Show that (13.18) implies asymptotic second-order self similarity (13.19). *Hint:* Use (13.14), and note that (13.18) is equivalent to $\mathsf{E}[Y_n^2]/n^{2H} \to \sigma_\infty^2$.

11. Show that a process is asymptotically second-order self similar; i.e., (13.19) holds, if and only if the conditions

$$\lim_{m\to\infty} \rho^{(m)}(n) \;=\; \frac{1}{2}[|n+1|^{2H} \Leftrightarrow 2|n|^{2H} + |n \Leftrightarrow 1|^{2H}],$$

and

$$\lim_{m\to\infty} \frac{C^{(m)}(0)}{m^{2H-2}} \;=\; \sigma_\infty^2$$

both hold.

12. Show that the covariance function corresponding to the power spectral density of Example 13.2 is

$$C(n) \;=\; \frac{(\Leftrightarrow 1)^n,\ (1 \Leftrightarrow 2d)}{,\ (n+1 \Leftrightarrow d),\ (1 \Leftrightarrow d \Leftrightarrow n)}.$$

This result is due to Hosking [23, Theorem 1(d)]. *Hints:* First show that

$$C(n) \;=\; \frac{1}{\pi}\int_0^\pi [4\sin^2(\nu/2)]^{-d}\cos(n\nu)\,d\nu.$$

Second, use the change-of-variable $\theta = 2\pi \Leftrightarrow \nu$ to show that

$$\frac{1}{\pi}\int_\pi^{2\pi} [4\sin^2(\nu/2)]^{-d}\cos(n\nu)\,d\nu \;=\; C(n).$$

Third, use the formula [16, p. 372]

$$\int_0^\pi \sin^{p-1}(t)\cos(at)\,dt \;=\; \frac{\pi\cos(a\pi/2),\ (p+1)2^{1-p}}{p,\ \left(\frac{p+a+1}{2}\right),\ \left(\frac{p \Leftrightarrow a+1}{2}\right)}.$$

13. Prove that $K_n(\alpha) \to K(\alpha)$ as follows. Put

$$J_n(\alpha) \;:=\; n^{\alpha} \int_{-1/2}^{1/2} \frac{1}{|f|^{1-\alpha}} \left[\frac{\sin(n\pi f)}{n\pi f}\right]^2 df.$$

First show that

$$J_n(\alpha) \;\to\; \pi^{-\alpha} \int_{-\infty}^{\infty} \frac{1}{|\theta|^{1-\alpha}} \left[\frac{\sin\theta}{\theta}\right]^2 d\theta.$$

Then show that $K_n(\alpha) \Leftrightarrow J_n(\alpha) \to 0$; take care to write this difference as an integral, and to break up the range of integration into $[0,\delta]$ and $[\delta, 1/2]$, where $0 < \delta < 1/2$ is chosen small enough that when $|f| < \delta$,

$$\left| \left[\frac{\pi f}{\sin(\pi f)}\right]^2 \Leftrightarrow 1 \right| \;<\; \varepsilon.$$

14. Show that for $0 < \alpha < 1$,

$$\int_0^{\infty} \theta^{\alpha-3} \sin^2\theta \, d\theta \;=\; \frac{2^{1-\alpha}\cos(\alpha\pi/2),\,(\alpha)}{(1 \Leftrightarrow \alpha)(2 \Leftrightarrow \alpha)}.$$

Remark. The formula actually holds for complex α with $0 < \mathrm{Re}\,\alpha < 2$ [16, p. 447].

Hints: (*i*) Fix $0 < \varepsilon < r < \infty$, and apply integration by parts to

$$\int_{\varepsilon}^{r} \theta^{\alpha-3} \sin^2\theta \, d\theta$$

with $u = \sin^2\theta$ and $dv = \theta^{\alpha-3}\,d\theta$.

(*ii*) Apply integration by parts to the integral

$$\int t^{\alpha-2} \sin t \, dt$$

with $u = \sin t$ and $dv = t^{\alpha-2} dt$.

(*iii*) Use the fact that for $0 < \alpha < 1$,[†]

$$\lim_{\substack{r\to\infty \\ \varepsilon\to 0}} \int_{\varepsilon}^{r} t^{\alpha-1} e^{-jt}\, dt \;=\; e^{-j\alpha\pi/2},\,(\alpha).$$

[†] For $s > 0$, a change of variable shows that

$$\lim_{\substack{r\to\infty \\ \varepsilon\to 0}} \int_{\varepsilon}^{r} t^{\alpha-1} e^{-st}\, dt \;=\; \frac{,\,(\alpha)}{s^{\alpha}}.$$

As in Notes 5 and 6 in Chapter 3, a permanence of form argument allows us to set $s = j = e^{j\pi/2}$.

15. Let $S(f)$ be given by (13.15). For $1/2 < H < 1$, put $\alpha = 2 \Leftrightarrow 2H$.

    (a) Evaluate the limit in (13.20). *Hint:* You may use the fact that

    $$\sum_{i=1}^{\infty} \frac{1}{|i+f|^{2H+1}}$$

    converges uniformly for $|f| \le 1/2$ and is therefore a continuous and bounded function.

    (b) Evaluate

    $$\int_{-1/2}^{1/2} S(f)\, df.$$

    Hint: The above integral is equal to $C(0) = \sigma^2$. Since (13.15) corresponds to a second-order self-similar process, not just an *asymptotically* second-order self-similar process, $\sigma^2 = \sigma_\infty^2$. Now apply (13.21).

Problems §13.4: Long-Range Dependence

16. Show directly that if a wide-sense stationary sequence has the covariance function $C(n)$ given in Problem 12, then the process is long-range dependent; i.e., (13.25) holds with appropriate values of α and c [23, Theorem 1(d)]. *Hints:* Use the Remark following Problem 11 in Chapter 3, Stirling's formula,

    $$, (x) \;\sim\; \sqrt{2\pi} x^{x-1/2} e^{-x},$$

    and the formula $(1 + d/n)^n \to e^d$.

17. For $0 < \alpha < 1$, show that

    $$\lim_{n\to\infty} \frac{\sum_{\nu=k}^{n-1} \nu^{-\alpha}}{n^{1-\alpha}} \;=\; \frac{1}{1 \Leftrightarrow \alpha}.$$

    Hints: Rewrite (13.30) in the form

    $$B_n + n^{-\alpha} \Leftrightarrow k^{-\alpha} \;\le\; I_n \;\le\; B_n.$$

    Then

    $$1 \;\le \frac{B_n}{I_n} \;\le\; 1 + \frac{k^{-\alpha} \Leftrightarrow n^{-\alpha}}{I_n}.$$

    Show that $I_n/n^{1-\alpha} \to 1/(1 \Leftrightarrow \alpha)$, and note that this implies $I_n/n^{-\alpha} \to \infty$.

18. For $0 < \alpha < 1$, show that

    $$\lim_{n\to\infty} \frac{\sum_{\nu=k}^{n-1} \nu^{1-\alpha}}{n^{2-\alpha}} \;=\; \frac{1}{2 \Leftrightarrow \alpha}.$$

Problems §13.5: ARMA Processes

19. Use the bound $|\alpha_n| \le M(1+\delta/2)^{-n}$ to show that $\sum_n |h_n| < \infty$, where

$$h_n \;=\; \sum_{k=\max(0,n-q)}^{n} \alpha_k b_{n-k}.$$

20. Assume (13.32) holds and that $A(z)$ has all roots strictly inside the unit circle. Show that (13.33) must hold. *Hint:* Compute the convolution

$$\sum_n \alpha_{m-n} Y_n$$

first for Y_n replaced by the left-hand side of (13.32) and again for Y_n replaced by the right-hand side of (13.32).

21. Let $\sum_{k=0}^{\infty} |h_k| < \infty$. Show that if X_n is WSS, then $Y_n = \sum_{k=0}^{\infty} h_k X_{n-k}$ and X_n are J-WSS.

Bibliography

[1] M. Abramowitz and I. A. Stegun, eds. *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 1970.

[2] J. Beran, *Statistics for Long-Memory Processes*. New York: Chapman & Hall, 1994.

[3] P. Billingsley, *Probability and Measure*. New York: Wiley, 1979.

[4] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[5] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York: Springer-Verlag, 1987.

[6] R. C. Buck, *Advanced Calculus*, 3rd ed. New York: McGraw-Hill, 1978.

[7] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–507, 1952.

[8] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 2nd ed. New York: Springer, 1988.

[9] R. V. Churchill, J. W. Brown, and R. F. Verhey, *Complex Variables and Applications*, 3rd ed. New York: McGraw-Hill, 1976.

[10] J. W. Craig, "A new, simple and exact result for calculating the probability of error for two-dimensional signal constellations," in *Proc. IEEE Milit. Commun. Conf. MILCOM '91*, McLean, VA, Oct. 1991, pp. 571–575.

[11] H. Cramér, "Sur un nouveaux théorème-limite de la théorie des probabilités," *Actualités Scientifiques et Industrielles 736*, pp. 5–23. *Colloque consacré à la théorie des probabilités*, vol. 3, Hermann, Paris, Oct. 1937.

[12] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," *Perf. Eval. Rev.*, vol. 24, pp. 160–169, 1996.

[13] M. H. A. Davis, *Linear Estimation and Stochastic Control*. London: Chapman and Hall, 1977.

[14] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed. New York: Wiley, 1968.

[15] L. Gonick and W. Smith, *The Cartoon Guide to Statistics*. New York: HarperPerennial, 1993.

[16] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. Orlando, FL: Academic Press, 1980.

[17] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford: Oxford University Press, 2001.

[18] S. Haykin and B. Van Veen, *Signals and Systems*. New York: Wiley, 1999.

[19] C. W. Helstrom, *Statistical Theory of Signal Detection*, 2nd ed. Oxford: Pergamon, 1968.

[20] C. W. Helstrom, *Probability and Stochastic Processes for Engineers*, 2nd ed. New York: Macmillan, 1991.

[21] P. G. Hoel, S. C. Port, and C. J. Stone, *Introduction to Probability Theory*. Boston: Houghton Mifflin, 1971.

[22] K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[23] J. R. M. Hosking, "Fractional differencing," *Biometrika*, vol. 68, no. 1, pp. 165–176, 1981.

[24] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. New York: Academic Press, 1975.

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

[26] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[27] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd ed. Reading, MA: Addison-Wesley, 1994.

[28] N. Likhanov, "Bounds on the buffer occupancy probability with self-similar traffic," pp. 193–213, in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. New York: Wiley, 2000.

[29] A. O'Hagen, *Probability: Methods and Measurement*. London: Chapman and Hall, 1988.

[30] A. V. Oppenheim and A. S. Willsky, with S. H. Nawab, *Signals & Systems*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1997.

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984.

[32] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226-244, 1995.

[33] B. Picinbono, *Random Signals and Systems*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[34] S. Ross, *A First Course in Probability*, 3rd ed. New York: Macmillan, 1988.

[35] R. I. Rothenberg, *Probability and Statistics*. San Diego: Harcourt Brace Jovanovich, 1991.

[36] G. Roussas, *A First Course in Mathematical Statistics*. Reading, MA: Addison-Wesley, 1973.

[37] H. L. Royden, *Real Analysis*, 2nd ed. New York: MacMillan, 1968.

[38] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[39] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. New York: Chapman & Hall, 1994.

[40] A. N. Shiryayev, *Probability*. New York: Springer, 1984.

[41] M. K. Simon, "A new twist on the Marcum Q function and its application," *IEEE Commun. Lett.*, vol. 2, no. 2, pp. 39–41, Feb. 1998.

[42] M. K. Simon and D. Divsalar, "Some new twists to problems involving the Gaussian probability integral," *IEEE Trans. Commun.*, vol. 46, no. 2, pp. 200–210, Feb. 1998.

[43] M. K. Simon and M.-S. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels," *Proc. IEEE*, vol. 86, no. 9, pp. 1860–1877, Sep. 1998.

[44] Ya. G. Sinai, "Self-similar probability distributions," *Theory of Probability and its Applications*, vol. XXI, no. 1, pp. 64–80, 1976.

[45] J. L. Snell, *Introduction to Probability*. New York: Random House, 1988.

[46] E. W. Stacy, "A generalization of the gamma distribution," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1187–1192, Sept. 1962.

[47] H. Stark and J. W. Woods, *Probability and Random Processes with Applications to*

Signal Processing, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[48] B. Tsybakov and N. D. Georganas, "Self-similar processes in communication networks," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1713–1725, Sept. 1998.

[49] Y. Viniotis, *Probability and Random Processes for Electrical Engineers*. Boston: WCB/McGraw-Hill, 1998.

[50] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*. New York: Springer, 1985.

[51] R. D. Yates and D. J. Goodman, *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*. New York: Wiley, 1999.

[52] R. E. Ziemer, *Elements of Engineering Probability and Statistics*. Upper Saddle River, NJ: Prentice Hall, 1997.

Index

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

Continuous Random Variables

---

uniform$[a, b]$

$f_X(x) = \dfrac{1}{b-a}$ and $F_X(x) = \dfrac{x-a}{b-a}$, $a \le x \le b$.

$\mathsf{E}[X] = \dfrac{a+b}{2}$, $\mathsf{var}(X) = \dfrac{(b-a)^2}{12}$, $M_X(s) = \dfrac{e^{sb} - e^{sa}}{s(b-a)}$.

---

exponential, exp(λ)

$f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$, $x \ge 0$.

$\mathsf{E}[X] = 1/\lambda$, $\mathsf{var}(X) = 1/\lambda^2$, $\mathsf{E}[X^n] = n!/\lambda^n$.

$M_X(s) = \lambda/(\lambda - s)$, $\mathrm{Re}\, s < \lambda$.

Laplace(λ)

$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|}$.

$\mathsf{E}[X] = 0$, $\mathsf{var}(X) = 2/\lambda^2$. $M_X(s) = \lambda^2/(\lambda^2 - s^2)$, $-\lambda < \mathrm{Re}\, s < \lambda$.

Gaussian or normal, $N(m, \sigma^2)$

$f_X(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-m}{\sigma}\right)^2\right]$. $F_X(x) = \texttt{normcdf(x,m,sigma)}$.

$\mathsf{E}[X] = m$, $\mathsf{var}(X) = \sigma^2$, $\mathsf{E}[(X-m)^{2n}] = 1 \cdot 3 \cdots (2n-3)(2n-1)\sigma^{2n}$,

$M_X(s) = e^{sm + s^2\sigma^2/2}$.

---

gamma(p, λ)

$f_X(x) = \lambda \dfrac{(\lambda x)^{p-1} e^{-\lambda x}}{,(p)}$, $x > 0$, where $,(p) := \int_0^\infty x^{p-1}\, e^{-x}\, dx$, $p > 0$.

Recall that $,(p) = (p-1) \cdot ,(p-1)$, $p > 1$.

$F_X(x) = \texttt{gamcdf(x,p,1/lambda)}$.

$\mathsf{E}[X^n] = \dfrac{,(n+p)}{\lambda^n ,(p)}$. $M_X(s) = \left(\dfrac{\lambda}{\lambda - s}\right)^p$, $\mathrm{Re}\, s < \lambda$.

Note that gamma$(1, \lambda)$ is the same as exp(λ).

Continuous Random Variables (continued)

Erlang(m, λ) := gamma(m, λ), m = integer

Since , $(m) = (m-1)!$

$$f_X(x) = \lambda \frac{(\lambda x)^{m-1}\, e^{-\lambda x}}{(m-1)!} \quad \text{and} \quad F_X(x) = 1 - \sum_{k=0}^{m-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x}, \quad x > 0.$$

Note that Erlang$(1, \lambda)$ is the same as exp(λ).

---

chi-squared(k) := gamma($k/2, 1/2$)

If k is an even integer, then chi-squared(k) is the same as Erlang$(k/2, 1/2)$.
Since , $(1/2) = \sqrt{\pi}$,

for $k = 1$, $f_X(x) = \dfrac{e^{-x/2}}{\sqrt{2\pi x}}, \quad x > 0.$

Since , $\left(\frac{2m+1}{2}\right) = \frac{(2m-1)\cdots 5\cdot 3\cdot 1}{2^m}\sqrt{\pi}$,

for $k = 2m+1$, $f_X(x) = \dfrac{x^{m-1/2} e^{-x/2}}{(2m-1)\cdots 5\cdot 3\cdot 1\sqrt{2\pi}}, \quad x > 0.$

$F_X(x) =$ `chi2cdf(x,k)`.

Note that chi-squared(2) is the same as exp$(1/2)$.

Rayleigh(λ)

$f_X(x) = \dfrac{x}{\lambda^2} e^{-(x/\lambda)^2/2}$ and $F_X(x) = 1 - e^{-(x/\lambda)^2/2}, \quad x \geq 0.$

$\mathsf{E}[X] = \lambda\sqrt{\pi/2}, \quad \mathsf{E}[X^2] = 2\lambda^2, \quad \mathsf{var}(X) = \lambda^2(2 - \pi/2).$

$\mathsf{E}[X^n] = 2^{n/2}\lambda^n,\ (1 + n/2).$

Weibull(p, λ)

$$f_X(x) = \lambda p x^{p-1} e^{-\lambda x^p} \quad \text{and} \quad F_X(x) = 1 - e^{-\lambda x^p}, \quad x > 0.$$

$$\mathsf{E}[X^n] = \frac{,\ (1 + n/p)}{\lambda^{n/p}}.$$

Note that Weibull$(2, \lambda)$ is the same as Rayleigh$\left(1/\sqrt{2\lambda}\right)$ and that Weibull$(1, \lambda)$ is the same as exp(λ).

Cauchy(λ)

$$f_X(x) = \frac{\lambda/\pi}{\lambda^2 + x^2}, \quad F_X(x) = \frac{1}{\pi}\tan^{-1}\left(\frac{x}{\lambda}\right) + \frac{1}{2}.$$

$\mathsf{E}[X] =$ undefined, $\quad \mathsf{E}[X^2] = \infty, \quad \varphi_X(\nu) = e^{-\lambda|\nu|}.$

Odd moments are not defined; even moments are infinite. Since the first moment is not defined, central moments, including the variance, are not defined.
