CHAPTER 7

Probabilistic Models

INTRODUCTION

In the preceding chapters our emphasis has been on deterministic signals. In the remainder of this text we expand the class of signals considered to include those that are based on probabilistic models, referred to as random or stochastic processes. In introducing this important class of signals, we begin in this chapter with a review of the basics of probability and random variables. We assume that you have encountered this foundational material in a previous course, but include a review here for convenient reference and to establish notation. In the following chapter and beyond, we apply these concepts to define and discuss the class of random signals.

7.1 THE BASIC PROBABILITY MODEL

Associated with a basic probability model are the following three components, as indicated in Figure 7.1:

1. **Sample Space** The sample space Ψ is the set of all possible outcomes ψ of the probabilistic experiment that the model represents. We require that one and only one outcome be produced in each experiment with the model.
2. **Event Algebra** An event algebra is a collection of subsets of the sample space — referred to as events in the sample space — chosen such that unions of events and complements of events are themselves events (i.e., are in the collection of subsets). We say that a particular event has occurred if the outcome of the experiment lies in this event subset; thus Ψ is the "certain event" because it always occurs, and the empty set $\emptyset$ is the "impossible event" because it never occurs. Note that intersections of events are also events, because intersections can be expressed in terms of unions and complements.
3. **Probability Measure** A probability measure associates with each event A a number $P(A)$, termed the probability of A, in such a way that:
   - **(a)** $P(A) \geq 0$;
   - **(b)** $P(\Psi) = 1$;
   - **(c)** If $A \cap B = \emptyset$, i.e., if events A and B are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B) \, .$$

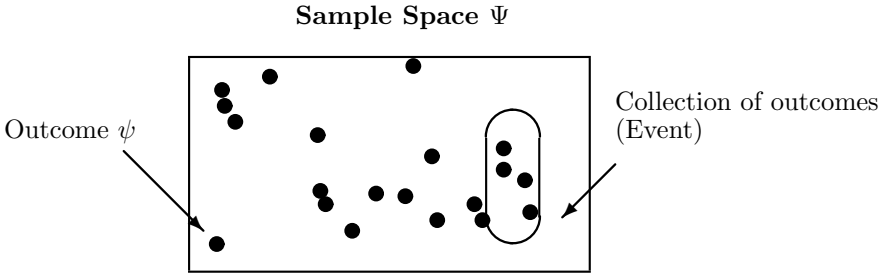

FIGURE 7.1 Sample space and events.

Note that for any particular case we often have a range of options in specifying what constitutes an outcome, in defining an event algebra, and in assigning a probability measure. It is generally convenient to have as few elements or outcomes as possible in a sample space, but we need enough of them to enable specification of the events of interest to us. It is typically convenient to pick the smallest event algebra that contains the events of interest. We also require that there be an assignment of probabilities to events that is consistent with the above conditions. This assignment may be made on the basis of symmetry arguments or in some other way that is suggested by the particular application.

7.2 CONDITIONAL PROBABILITY, BAYES' RULE, AND INDEPENDENCE

The probability of event A, given that event B has occurred, is denoted by $P(A|B)$. Knowing that B has occurred in effect reduces the sample space to the outcomes in B, so a natural definition of the conditional probability is

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \text{ if } P(B) > 0 \, . \tag{7.1}$$

It is straightforward to verify that this definition of conditional probability yields a valid probability measure on the sample space B. The preceding equation can also be rearranged to the form

$$P(A \cap B) = P(A|B)P(B) \, . \tag{7.2}$$

We often write $P(AB)$ or $P(A, B)$ for the joint probability $P(A \cap B)$. If $P(B) = 0$, then the conditional probability in (7.1) is undefined.

By symmetry, we can also write

$$P(A \cap B) = P(B|A)P(A) \tag{7.3}$$

Combining the preceding two equations, we obtain one form of Bayes' rule (or theorem), which is at the heart of much of what we'll do with signal detection,

classification, and estimation:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{7.4}$$

A more detailed form of Bayes' rule can be written for the conditional probability of one of a set of events $\{B_j\}$ that are mutually exclusive and collectively exhaustive, i.e. $B_\ell \cap B_m = \emptyset$ if $\ell \neq m$, and $\bigcup_j B_j = \Psi$. In this case,

$$P(A) = \sum_j P(A \cap B_j) = \sum_j P(A|B_j)P(B_j) \tag{7.5}$$

so that

$$P(B_\ell|A) = \frac{P(A|B_\ell)P(B_\ell)}{\sum_j P(A|B_j)P(B_j)} \tag{7.6}$$

Events A and B are said to be independent if

$$P(A|B) = P(A) \tag{7.7}$$

or equivalently if the joint probability factors as

$$P(A \cap B) = P(A)P(B)\ . \tag{7.8}$$

More generally, a collection of events is said to be mutually independent if the probability of the intersection of events from this collection, taken any number at a time, is always the product of the individual probabilities. Note that pairwise independence is not enough. Also, two sets of events $\mathcal{A}$ and $\mathcal{B}$ are said to be independent of each other if the probability of an intersection of events taken from these two sets always factors into the product of the joint probability of those events that are in $\mathcal{A}$ and the joint probability of those events that are in $\mathcal{B}$.

EXAMPLE 7.1 Transmission errors in a communication system

A communication system transmits symbols labeled A, B, and C. Because of errors (noise) introduced by the channel, there is a nonzero probability that for each transmitted symbol, the received symbol differs from the transmitted one. Table 7.1 describes the joint probability for each possible pair of transmitted and received symbols under a certain set of system conditions.

	Symbol received		
Symbol sent	A	B	C
A	0.05	0.10	0.09
B	0.13	0.08	0.21
C	0.12	0.07	0.15

TABLE 7.1 Joint probability for each possible pair of transmitted and received symbols

For notational convenience let's use A_s, B_s, C_s to denote the events that A, B or C respectively is sent, and A_r, B_r, C_r to denote A, B or C respectively being received. So, for example, $P(A_r, B_s) = 0.13$ and $P(C_r, C_s) = 0.15$. To determine the marginal probability $P(A_r)$, we sum the probabilities for all the mutually exclusive ways that A is received. So, for example,

$$\begin{aligned} P(A_r) &= P(A_r, A_s) + P(A_r, B_s) + P(A_r, C_s) \\ &= .05 + .13 + .12 = 0.3 \, . \end{aligned} \tag{7.9}$$

Similarly we can determine the marginal probability $P(A_s)$ as

$$P(A_s) = P(A_r, A_s) + P(B_r, A_s) + P(C_r, A_s) = 0.24 \tag{7.10}$$

In a communication context, it may be important to know the probability, for example, that C was sent, given that B was received, i.e., $P(C_s|B_r)$. That information is not entered directly in the table but can be calculated from it using Bayes' rule. Specifically, the desired conditional probability can be expressed as

$$P(C_s|B_r) = \frac{P(C_s, B_r)}{P(B_r)} \tag{7.11}$$

The numerator in (7.11) is given directly in the table as .07. The denominator is calculated as $P(B_r) = P(B_r, A_s) + P(B_r, B_s) + P(B_r, C_s) = 0.25$. The result then is that $P(C_s|B_r) = 0.28$.

In communication systems it is also often of interest to measure or calculate the probability of a transmission error. Denoting this by P_t it would correspond to any of the following mutually exclusive events happening:

$$(A_s \cap B_r), (A_s \cap C_r), (B_s \cap A_r), (B_s \cap C_r), (C_s \cap A_r), (C_s \cap B_r) \tag{7.12}$$

P_t is therefore the sum of the probabilities of these six mutually exclusive events, and all these probabilities can be read directly from the table in the off-diagonal locations, yielding $P_t = 0.72$.

7.3 RANDOM VARIABLES

A real-valued random variable $X(\,\cdot\,)$ is a function that maps each outcome ψ of a probabilistic experiment to a real number $X(\psi)$, which is termed the *realization* of (or value taken by) the random variable in that experiment. An additional technical requirement imposed on this function is that the set of outcomes $\{\psi\}$ that maps to the interval $X \leq x$ must be an event in Ψ, for all real numbers x. We shall typically just write the random variable as X instead of $X(\,\cdot\,)$ or $X(\psi)$.

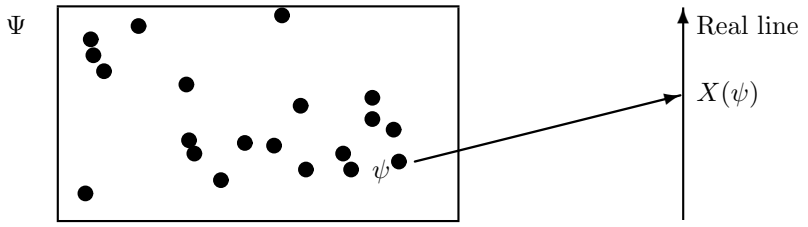

FIGURE 7.2 A random variable.

It is often also convenient to consider random variables taking values that are not specified as real numbers but rather a finite or countable set of labels, say $L_0, L_1, L_2, \ldots$. For instance, the random status of a machine may be tracked using the labels Idle, Busy, and Failed. Similarly, the random presence of a target in a radar scan can be tracked using the labels Absent and Present. We can think of these labels as comprising a set of mutually exclusive and collectively exhaustive events, where each such event comprises all the outcomes that carry that label. We refer to such random variables as random events, mapping each outcome ψ of a probabilistic experiment to the label $L(\psi)$, chosen from the possible values $L_0, L_1, L_2, \ldots$. We shall typically just write L instead of $L(\psi)$.

7.4 CUMULATIVE DISTRIBUTION, PROBABILITY DENSITY, AND PROBABILITY MASS FUNCTION FOR RANDOM VARIABLES

Cumulative Distribution Functions For a (real-valued) random variable X, the probability of the event comprising all ψ for which $X(\psi) \leq x$ is described using the cumulative distribution function (CDF) $F_X(x)$:

$$F_X(x) = P(X \leq x) \, . \tag{7.13}$$

We can therefore write

$$P(a < X \leq b) = F_X(b) - F_X(a) \, . \tag{7.14}$$

In particular, if there is a nonzero probability that X takes a specific value x_1, i.e. if $P(X = x_1) > 0$, then $F_X(x)$ will have a jump at x_1 of height $P(X = x_1)$, and $F_X(x_1) - F_X(x_1-) = P(X = x_1)$. The CDF is nondecreasing as a function of x; it starts from $F_X(-\infty) = 0$ and rises to $F_X(\infty) = 1$.

A related function is the conditional CDF $F_{X|L}(x|L_i)$, used to describe the distribution of X conditioned on some random event L taking the specific value L_i, and assuming $P(L = L_i) > 0$:

$$F_{X|L}(x|L_i) = P(X \leq x|L = L_i) = \frac{P(X \leq x, L = L_i)}{P(L = L_i)} \, . \tag{7.15}$$

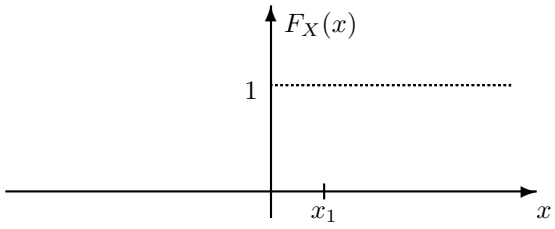

FIGURE 7.3 Example of a CDF.

Probability Density Functions The probability density function (PDF) $f_X(x)$ of the random variable X is the derivative of $F_X(x)$:

$$f_X(x) = \frac{dF_X(x)}{dx} \ . \tag{7.16}$$

It is of course always non-negative because $F_X(x)$ is nondecreasing. At points of discontinuity in $F_X(x)$, corresponding to values of x that have non-zero probability of occurring, there will be (Dirac) impulses in $f_X(x)$, of strength or area equal to the height of the discontinuity. We can write

$$P(a < X \leq b) = \int_a^b f_X(x)\, dx \ . \tag{7.17}$$

(Any impulse of $f_X(x)$ at b would be included in the integral, while any impulse at a would be left out — i.e. the integral actually goes from $a+$ to $b+$.) We can heuristically think of $f_X(x)\, dx$ as giving the probability that X lies in the interval $(x - dx, x]$:

$$P(x - dx < X \leq x) \approx f_X(x)\, dx \ . \tag{7.18}$$

Note that at values of x where $f_X(x)$ does not have an impulse, the probability of X having the value x is zero, i.e., $P(X = x) = 0$.

A related function is the conditional PDF $f_{X|L}(x|L_i)$, defined as the derivative of $F_{X|L}(x|L_i)$ with respect to x.

Probability Mass Function A real-valued discrete random variable X is one that takes only a finite or countable set of real values, $\{x_1, x_2, \cdots\}$. (Hence this is actually a random event — as defined earlier — but specified numerically rather than via labels.) The CDF in this case would be a "staircase" function, while the PDF would be zero everywhere, except for impulses at the x_j, with strengths corresponding to the respective probabilities of the x_j. These strengths/probabilities are conveniently described by the probability mass function (PMF) $p_X(x)$, which gives the probability of the event $X = x_j$:

$$P(X = x_j) = p_X(x_j) \ . \tag{7.19}$$

7.5 JOINTLY DISTRIBUTED RANDOM VARIABLES

We almost always use models involving multiple (or compound) random variables. Such situations are described by joint probabilities. For example, the joint CDF of two random variables X and Y is

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \, . \tag{7.20}$$

The corresponding joint PDF is

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \, \partial y} \tag{7.21}$$

and has the heuristic interpretation that

$$P(x - dx < X \leq x \, , y - dy < Y \leq y) \approx f_{X,Y}(x,y) \, dx \, dy \, . \tag{7.22}$$

The marginal PDF $f_X(x)$ is defined as the PDF of the random variable X considered on its own, and is related to the joint density $f_{X,Y}(x,y)$ by

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) \, dy \, . \tag{7.23}$$

A similar expression holds for the marginal PDF $f_Y(y)$.

We have already noted that when the model involves a random variable X and a random event L, we may work with the conditional CDF

$$F_{X|L}(x|L_i) = P(X \leq x | L = L_i) = \frac{P(X \leq x, L = L_i)}{P(L = L_i)} \, , \tag{7.24}$$

provided $P(L = L_i) > 0$. The derivative of this function with respect to x gives the conditional PDF $f_{X|L}(x|L_i)$. When the model involves two continuous random variables X and Y, the corresponding function of interest is the conditional PDF $f_{X|Y}(x|y)$ that describes the distribution of X, given that $Y = y$. However, for a continuous random variable Y, $P(Y = y) = 0$, so even though the following definition may seem natural, its justification is more subtle:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \, . \tag{7.25}$$

To see the plausibility of this definition, note that the conditional PDF $f_{X|Y}(x|y)$ must have the property that

$$f_{X|Y}(x|y) \, dx \approx P(x - dx < X \leq x \mid y - dy < Y \leq y) \tag{7.26}$$

but by Bayes' rule the quantity on the right in the above equation can be rewritten as

$$P(x - dx < X \leq x \mid y - dy < Y \leq y) \approx \frac{f_{X,Y}(x,y) \, dx \, dy}{f_Y(y) dy} \, . \tag{7.27}$$

Combining the latter two expressions yields the definition of $f_{X|Y}(x|y)$ given in (7.25).

Using similar reasoning, we can obtain relationships such as the following:

$$P(L = L_i|X = x) = \frac{f_{X|L}(x|L_i)P(L = L_i)}{f_X(x)} \ . \tag{7.28}$$

Two random variables X and Y are said to be independent or statistically independent if their joint PDF (or equivalently their joint CDF) factors into the product of the individual ones:

$$\begin{array}{rcll} f_{X,Y}(x,y) & = & f_X(x)f_Y(y) \ , & \text{or} \\ F_{X,Y}(x,y) & = & F_X(x)F_Y(y) \ . & \end{array} \tag{7.29}$$

This condition turns out to be equivalent to having any collection of events defined in terms of X be independent of any collection of events defined in terms of Y.

For a set of more than two random variables to be independent, we require that the joint PDF (or CDF) of random variables from this set factors into the product of the individual PDFs (respectively, CDFs). One can similarly define independence of random variables and random events.

EXAMPLE 7.2 Independence of events

To illustrate some of the above definitions and concepts in the context of random variables and random events, consider two independent random variables X and Y for which the marginal PDFs are uniform between zero and one:

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Because X and Y are independent, the joint PDF $f_{X,Y}(x,y)$ is given by

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

We define the events A, B, C and D as follows:

$$A = \left\{y > \frac{1}{2}\right\}, \quad B = \left\{y < \frac{1}{2}\right\}, \quad C = \left\{x < \frac{1}{2}\right\},$$
$$D = \left\{x < \frac{1}{2} \text{ and } y < \frac{1}{2}\right\} \cup \left\{x > \frac{1}{2} \text{ and } y > \frac{1}{2}\right\}.$$

These events are illustrated pictorially in Figure 7.4

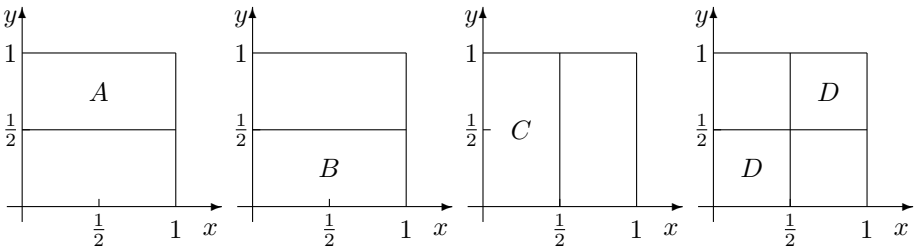

FIGURE 7.4 Illustration of events A, B, C, and D, for Example 7.2

Questions that we might ask include whether these events are pairwise independent, e.g. whether A and C are independent. To answer such questions, we consider whether the joint probability factors into the product of the individual probabilities. So, for example,

$$P(A \cap C) = P\left(y > \frac{1}{2}, x < \frac{1}{2}\right) = \frac{1}{4}$$

$$P(A) = P(C) = \frac{1}{2}$$

Since $P(A \cap C) = P(A)P(C)$, events A and C are independent. However,

$$P(A \cap B) = P\left(y > \frac{1}{2}, y < \frac{1}{2}\right) = 0$$

$$P(A) = P(B) = \frac{1}{2}$$

Since $P(A \cap B) \neq P(A)P(B)$, events A and B are not independent.

Note that $P(A \cap C \cap D) = 0$ since there is no region where all three sets overlap. However, $P(A) = P(C) = P(D) = \frac{1}{2}$, so $P(A \cap C \cap D) \neq P(A)P(C)P(D)$ and the events A, C, and D are not mutually independent, even though they are easily seen to be pairwise independent. For a collection of events to be independent, we require the probability of the intersection of any of the events to equal the product of the probabilities of each individual event. So for the 3–event case, pairwise independence is a necessary but not sufficient condition for independence.

7.6 EXPECTATIONS, MOMENTS AND VARIANCE

For many purposes it suffices to have a more aggregated or approximate description than the PDF provides. The expectation — also termed the expected or mean or average value, or the first-moment — of the real-valued random variable X is

denoted by $E[X]$ or $\overline{X}$ or μ_X, and defined as

$$E[X] = \overline{X} = \mu_X = \int_{-\infty}^{\infty} x f_X(x)\, dx \ . \tag{7.30}$$

In terms of the probability "mass" on the real line, the expectation gives the location of the center of mass. Note that the expected value of a sum of random variables is just the sum of the individual expected values:

$$E[X+Y] = E[X] + E[Y] \ . \tag{7.31}$$

Other simple measures of where the PDF is centered or concentrated are provided by the median, which is the value of x for which $F_X(x) = 0.5$, and by the mode, which is the value of x for which $f_X(x)$ is maximum (in degenerate cases one or both of these may not be unique).

The variance or centered second-moment of the random variable X is denoted by σ_X^2 and defined as

$$\begin{aligned} \sigma_X^2 &= E[(X-\mu_X)^2] = \text{expected squared deviation from the mean} \\ &= \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x) dx \\ &= E[X^2] - \mu_X^2 \ , \end{aligned} \tag{7.32}$$

where the last equation follows on writing $(X-\mu_X)^2 = X^2 - 2\mu_X X + \mu_X^2$ and taking the expectation term by term. We refer to $E[X^2]$ as the second-moment of X. The square root of the variance, termed the standard deviation, is a widely used measure of the spread of the PDF.

The focus of many engineering models that involve random variables is primarily on the means and variances of the random variables. In some cases this is because the detailed PDFs are hard to determine or represent or work with. In other cases, the reason for this focus is that the means and variances completely determine the PDFs, as with the Gaussian (or normal) and uniform PDFs.

EXAMPLE 7.3 Gaussian and uniform random variables

Two common PDF's that we will work with are the Gaussian (or normal) density and the uniform density:

$$\begin{aligned} &\text{Gaussian:} \quad f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2} \\ &\text{Uniform:} \quad f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{7.33}$$

The two parameters m and σ that define the Gaussian PDF can be shown to be its mean and standard deviation respectively. Similarly, though the uniform density can be simply parametrized by its lower and upper limits a and b as above, an

equivalent parametrization is via its mean $m = (a+b)/2$ and standard deviation $\sigma = \sqrt{(b-a)^2/12}$.

There are useful statements that can be made for general PDFs on the basis of just the mean and variance. The most familiar of these is the Chebyshev inequality:

$$P\Big(\frac{|X-\mu_X|}{\sigma_X} \geq k\Big) \leq \frac{1}{k^2} \; . \tag{7.34}$$

This inequality implies that, for any random variable, the probability it lies at or more than 3 standard deviations away from the mean (on either side of the mean) is not greater than $(1/3^2) = 0.11$. Of course, for particular PDFs, much more precise statements can be made, and conclusions derived from the Chebyshev inequality can be very conservative. For instance, in the case of a Gaussian PDF, the probability of being more than 3 standard deviations away from the mean is only 0.0026, while for a uniform PDF the probability of being more than even 2 standard deviations away from the mean is precisely 0.

For much of our discussion we shall make do with evaluating the means and variances of the random variables involved in our models. Also, we will be highlighting problems whose solution only requires knowledge of means and variances.

The conditional expectation of the random variable X, given that the random variable Y takes the value y, is the real number

$$E[X|Y=y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx = g(y) \; , \tag{7.35}$$

i.e., this conditional expectation takes some value $g(y)$ when $Y = y$. We may also consider the random variable $g(Y)$, namely the function of the random variable Y that, for each $Y = y$, evaluates to the conditional expectation $E[X|Y=y]$. We refer to this random variable $g(Y)$ as the conditional expectation of X "given Y" (as opposed to "given $Y = y$"), and denote $g(Y)$ by $E[X|Y]$. Note that the expectation $E[g(Y)]$ of the random variable $g(Y)$, i.e. the iterated expectation $E[E[X|Y]]$, is well defined. What we show in the next paragraph is that this iterated expectation works out to something simple, namely $E[X]$. This result will be of particular use in the next chapter.

Consider first how to compute $E[X]$ when we have the joint PDF $f_{X,Y}(x,y)$. One way is to evaluate the marginal density $f_X(x)$ of X, and then use the definition of expectation in (7.30):

$$E[X] = \int_{-\infty}^{\infty} x\Big(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\Big)\,dx \; . \tag{7.36}$$

However, it is often simpler to compute the conditional expectation of X, given $Y = y$, then average this conditional expectation over the possible values of Y, using the marginal density of Y. To derive this more precisely, recall that

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) \tag{7.37}$$

and use this in (7.36) to deduce that

$$E[X] = \int_{-\infty}^{\infty} f_Y(y)\Big(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx\Big)\,dy = E_Y[E_{X|Y}[X|Y]]\ . \tag{7.38}$$

We have used subscripts on the preceding expectations in order to make explicit which densities are involved in computing each of them. More simply, one writes

$$E[X] = E[E[X|Y]]\ . \tag{7.39}$$

The preceding result has an important implication for the computation of the expectation of a function of a random variable. Suppose $X = h(Y)$, then $E[X|Y] = h(Y)$, so

$$E[X] = E[E[X|Y]] = \int_{-\infty}^{\infty} h(y) f_Y(y) dy\ . \tag{7.40}$$

This shows that we only need $f_Y(y)$ to calculate the expectation of a function of Y; to compute the expectation of $X = h(Y)$, we do not need to determine $f_X(x)$.

Similarly, if X is a function of *two* random variables, $X = h(Y, Z)$, then

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y,z) f_{Y,Z}(y,z) dy\,dz\ . \tag{7.41}$$

It is easy to show from this that if Y and Z are independent, and if $h(y,z) = g(y)\ell(z)$, then

$$E[g(Y)\ell(Z)] = E[g(Y)]E[\ell(Z)]\ . \tag{7.42}$$

7.7 CORRELATION AND COVARIANCE FOR BIVARIATE RANDOM VARIABLES

Consider a pair of jointly distributed random variables X and Y. Their marginal PDFs are simply obtained by projecting the probability mass along the y-axis and x-axis directions respectively:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\ , \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\ . \tag{7.43}$$

In other words, the PDF of X is obtained by integrating the joint PDF over all possible values of the other random variable Y — and similarly for the PDF of Y.

It is of interest, just as in the single-variable case, to be able to capture the location and spread of the bivariate PDF in some aggregate or approximate way, without having to describe the full PDF. And again we turn to notions of mean and variance. The mean value of the bivariate PDF is specified by giving the mean values of each of its two component random variables: the mean value has an x component that is $E[X]$, and a y component that is $E[Y]$, and these two numbers can be evaluated from the respective marginal densities. The center of mass of the bivariate PDF is thus located at

$$(x, y) = (E[X], E[Y])\ . \tag{7.44}$$

A measure of the spread of the bivariate PDF in the x direction may be obtained from the standard deviation σ_X of X, computed from $f_X(x)$; and a measure of the spread in the y direction may be obtained from σ_Y, computed similarly from $f_Y(y)$. However, these two numbers clearly only offer a partial view. We would really like to know what the spread is in a general direction rather than just along the two coordinate axes. We can consider, for instance, the standard deviation (or, equivalently, the variance) of the random variable Z defined as

$$Z = \alpha X + \beta Y \tag{7.45}$$

for arbitrary constants α and β. Note that by choosing α and β appropriately, we get $Z = X$ or $Z = Y$, and therefore recover the special coordinate directions that we have already considered; but being able to analyze the behavior of Z for arbitary α and β allows us to specify the behavior in all directions.

To visualize how Z behaves, note that $Z = 0$ when $\alpha x+\beta y = 0$. This is the equation of a straight line through the origin in the (x, y) plane, a line that indicates the precise combinations of values x and y that contribute to determining $f_Z(0)$, by projection of $f_{X,Y}(x, y)$ along the line. Let us call this the reference line. If Z now takes a nonzero value z, the corresponding set of (x, y) values lies on a line offset from but parallel to the reference line. We project $f_{X,Y}(x, y)$ along this new offset line to determine $f_Z(z)$.

Before seeing what computations are involved in determining the variance of Z, note that the mean of Z is easily found in terms of quantities we have already computed, namely $E[X]$ and $E[Y]$:

$$E[Z] = \alpha E[X] + \beta E[Y] \, . \tag{7.46}$$

As for the variance of Z, it is easy to establish from (7.45) and (7.46) that

$$\sigma_Z^2 = E[Z^2] - (E[Z])^2 = \alpha^2\sigma_X^2 + \beta^2\sigma_Y^2 + 2\alpha\beta\,\sigma_{X,Y} \tag{7.47}$$

where σ_X^2 and σ_Y^2 are the variances already computed along the coordinate directions x and y, and $\sigma_{X,Y}$ is the covariance of X and Y, also denoted by $\mathrm{cov}(X, Y)$ or $C_{X,Y}$, and defined as

$$\sigma_{X,Y} = \mathrm{cov}(X, Y) = C_{X,Y} = E[(X - E[X])(Y - E[Y])] \tag{7.48}$$

or equivalently

$$\sigma_{X,Y} = E[XY] - E[X]E[Y] \, . \tag{7.49}$$

where (7.49) follows from multiplying out the terms in parentheses in (7.48) and then taking term-by-term expectations. Note that when $Y = X$ we recover the familiar expressions for the variance of X. The quantity $E[XY]$ that appears in (7.49), i.e., the expectation of the product of the random variables, is referred to as the correlation or second cross-moment of X and Y (to distinguish it from the second self-moments $E[X^2]$ and $E[Y^2]$), and will be denoted by $R_{X,Y}$:

$$R_{X,Y} = E[XY] \, . \tag{7.50}$$

It is reassuring to note from (7.47) that the covariance $\sigma_{X,Y}$ is the only new quantity needed when going from mean and spread computations along the coordinate axes to such computations along any axis; we do not need a new quantity for each new direction. In summary, we can express the location of $f_{X,Y}(x,y)$ in an aggregate or approximate way in terms of the 1st-moments, $E[X]$, $E[Y]$; and we can express the spread around this location in an aggregate or approximate way in terms of the (central) 2nd-moments, σ_X^2 , σ_Y^2 , $\sigma_{X,Y}$.

It is common to work with a normalized form of the covariance, namely the correlation coefficient $\rho_{X,Y}$:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \ . \tag{7.51}$$

This normalization ensures that the correlation coefficient is unchanged if X and/or Y is multiplied by any nonzero constant or has any constant added to it. For instance, the centered and normalized random variables

$$V = \frac{X - \mu_X}{\sigma_X} \ , \qquad W = \frac{Y - \mu_Y}{\sigma_Y} \ , \tag{7.52}$$

each of which has mean 0 and variance 1, have the same correlation coefficient as X and Y. The correlation coefficient might have been better called the covariance coefficient, since it is defined in terms of the covariance and not the correlation of the two random variables, but this more helpful name is not generally utilized.

Invoking the fact that σ_Z^2 in (7.47) must be non-negative, and further noting from this equation that σ_Z^2/β^2 is quadratic in α, it can be proved by elementary analysis of the quadratic expression that

$$|\rho_{X,Y}| \leq 1 \ . \tag{7.53}$$

From the various preceding definitions, a positive correlation $R_{X,Y} > 0$ suggests that X and Y tend to take the same sign, on average, whereas a positive covariance $\sigma_{X,Y} > 0$ — or equivalently a positive correlation coefficient $\rho_{X,Y} > 0$ — suggests that the deviations of X and Y from their respective means tend to take the same sign, on average. Conversely, a negative correlation suggests that X and Y tend to take opposite signs, on average, while a negative covariance or correlation coefficient suggests that the deviations of X and Y from their means tend to take opposite signs, on average.

Since the correlation coefficient of X and Y captures some features of the relation between their deviations from their respective means, we might expect that the correlation coefficient can play a role in constructing an estimate of Y from measurements of X, or vice versa. We shall see in the next chapter, where linear minimum mean-square error (LMMSE) estimation is studied, that this is indeed the case.

The random variables X and Y are said to be uncorrelated (or linearly independent, a less common and potentially misleading term) if

$$E[XY] = E[X]E[Y] \ , \tag{7.54}$$

or equivalently if

$$\sigma_{X,Y} = 0 \quad \text{or} \quad \rho_{X,Y} = 0 \, . \tag{7.55}$$

Thus uncorrelated does not mean zero correlation (unless one of the random variables has an expected value of zero). Rather, uncorrelated means zero covariance. Again, a better term for uncorrelated might have been non-covariant, but this term is not widely used.

Note also that independent random variables X and Y, i.e., those for which

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \, , \tag{7.56}$$

are always uncorrelated, but the converse is not generally true: uncorrelated random variables may not be independent. If X and Y are independent, then $E[XY] = E[X]E[Y]$ so X and Y are uncorrelated. The converse does *not* hold in general. For instance, consider the case where the combination (X, Y) takes only the values $(1, 0)$, $(-1, 0)$. $(0, 1)$ and $(0, -1)$, each with equal probability $\frac{1}{4}$. Then X and Y are easily seen to be uncorrelated but dependent, i.e., not independent.

A final bit of terminology that we will shortly motivate and find useful occurs in the following definition: Two random variables X and Y are **orthogonal** if $E[XY] = 0$.

EXAMPLE 7.4 Perfect correlation, zero correlation

Consider the degenerate case where Y is given by a deterministic linear function of a random variable X (so Y is also a random variable, of course):

$$Y = \xi X + \zeta \, , \tag{7.57}$$

where ξ and ζ are constants. Then it is easy to show that $\rho_{X,Y} = 1$ if $\xi > 0$ and $\rho = -1$ if $\xi < 0$. Note that in this case the probability mass is entirely concentrated on the line defined by the above equation, so the bivariate PDF — if we insist on talking about it! — is a two-dimensional impulse (but this fact is not important in evaluating $\rho_{X,Y}$).

You should also have no difficulty establishing that $\rho_{X,Y} = 0$ if

$$Y = \xi X^2 + \zeta \tag{7.58}$$

and X has a PDF $f_X(x)$ that is even about 0, i.e., $f_X(-x) = f_X(x)$.

EXAMPLE 7.5 Bivariate Gaussian density

The random variables X and Y are said to be bivariate Gaussian or bivariate normal if their joint PDF is given by

$$f_{X,Y}(x,y) = c \, \exp\Big\{ -q\Big(\frac{x - \mu_X}{\sigma_X}, \frac{y - \mu_Y}{\sigma_Y} \Big) \Big\} \tag{7.59}$$

where c is a normalizing constant (so that the PDF integrates to 1) and $q(v,w)$ is a quadratic function of its two arguments v and w, expressed in terms of the correlation coefficient ρ of X and Y:

$$c = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \tag{7.60}$$

$$q(v,w) = \frac{1}{2(1-\rho^2)}(v^2 - 2\rho vw + w^2) \tag{7.61}$$

This density is the natural bivariate generalization of the familiar Gaussian density, and has several nice properties:

- The marginal densities of X and Y are Gaussian.
- The conditional density of Y, given $X = x$, is Gaussian with mean ρx and variance $\sigma_Y^2(1-\rho^2)$ (which evidently does not depend on the value of x); and similary for the conditional density of X, given $Y = y$.
- If X and Y are uncorrelated, i.e., if $\rho = 0$, then X and Y are actually independent, a fact that is not generally true for other bivariate random variables, as noted above.
- Any two affine (i.e., linear plus constant) combinations of X and Y are themselves bivariate Gaussian (e.g., $Q = X + 3Y + 2$ and $R = 7X + Y - 3$ are bivariate Gaussian).

The bivariate Gaussian PDF and indeed the associated notion of correlation were essentially discovered by the statistician Francis Galton (a first-cousin of Charles Darwin) in 1886, with help from the mathematician Hamilton Dickson. Galton was actually studying the joint distribution of the heights of parents and children, and found that the marginals and conditionals were well represented as Gaussians. His question to Dickson was: what joint PDF has Gaussian marginals and conditionals? The answer: the bivariate Gaussian! It turns out that there is a 2-dimensional version of the central limit theorem, with the bivariate Gaussian as the limiting density, so this is a reasonable model for two jointly distributed random variables in many settings. There are also natural generalization to many variables.

Some of the generalizations of the preceding discussion from two random variables to many random variables are fairly evident. In particular, the mean of a joint PDF

$$f_{X_1,X_2,\cdots,X_\ell}(x_1,x_2,\cdots,x_\ell) \tag{7.62}$$

in the ℓ-dimensional space of possible values has coordinates that are the respective individual means, $E[X_1],\cdots,E[X_\ell]$. The spreads in the coordinate directions are deduced from the individual (marginal) spreads, $\sigma_{X_1},\cdots,\sigma_{X_\ell}$. To be able to compute the spreads in *arbitrary* directions, we need all the additional $\ell(\ell-1)/2$ central 2nd moments, namely σ_{X_i,X_j} for all $1 \le i < j \le \ell$ (note that $\sigma_{X_j,X_i} = \sigma_{X_i,X_j}$) — but nothing more.

7.8 A VECTOR-SPACE PICTURE FOR CORRELATION PROPERTIES OF RANDOM VARIABLES

A vector-space picture is often useful as an aid to recalling the second-moment relationships between two random variables X and Y. This picture is not just a mnemonic: there is a very precise sense in which random variables can be thought of (or are) vectors in a vector space (of infinite dimensions), as long as we are only interested in their second-moment properties. Although we shall not develop this correspondence in any depth, it can be very helpful in conjecturing or checking answers in the linear minimum mean-square-error (LMMSE) estimation problems that we shall treat.

To develop this picture, we represent the random variables X and Y as vectors $\mathbf{X}$ and $\mathbf{Y}$ in some abstract vector space. For the squared lengths of these vectors, we take the second-moments of the associated random variables, $E[X^2]$ and $E[Y^2]$ respectively. Recall that in Euclidean vector space the squared length of a vector is the inner product of the vector with itself. This suggests that perhaps in our vector-space interpretation the inner product $< \mathbf{X}, \mathbf{Y} >$ between two general vectors $\mathbf{X}$ and $\mathbf{Y}$ should be defined as the correlation (or second cross-moment) of the associate random variables:

$$< \mathbf{X}, \mathbf{Y} >= E[XY] = R_{X,Y} \ . \tag{7.63}$$

This indeed turns out to be the definition that's needed. With this definition, the standard properties required of an inner product in a vector space are satisfied, namely:

Symmetry: $< \mathbf{X}, \mathbf{Y} >=< \mathbf{Y}, \mathbf{X} > \ .$

Linearity: $< \mathbf{X}, a_1\mathbf{Y_1} + a_2\mathbf{Y_2} >= a_1 < \mathbf{X}, \mathbf{Y_1} > + a_2 < \mathbf{X}, \mathbf{Y_2} >$

Positivity: $< \mathbf{X}, \mathbf{X} >$ is positive for $\mathbf{X} \neq \mathbf{0}$, and $\mathbf{0}$ otherwise.

This definition of inner product is also consistent with the fact that we often refer to two random variables as orthogonal when $E[XY] = 0$.

The centered random variables $X - \mu_X$ and $Y - \mu_Y$ can similary be represented as vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ in this abstract vector space, with squared lengths that are now the variances of the random variables X and Y:

$$\sigma_X^2 = E[(X - \mu_X)^2] \ , \qquad \sigma_Y^2 = E[(Y - \mu_Y)^2] \tag{7.64}$$

respectively. The lengths are therefore the standard deviations of the associated random variables, σ_X and σ_Y respectively. The inner product of the vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ becomes

$$< \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}} >= E[(X - \mu_X)(Y - \mu_Y)] = \sigma_{X,Y} \ , \tag{7.65}$$

namely the covariance of the random variables.

In Euclidean space the inner product of two vectors is given by the product of the lengths of the individual vectors and the cosine of the angle between them:

$$< \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}} >= \sigma_{X,Y} = \sigma_X \sigma_Y \cos(\theta) \ , \tag{7.66}$$

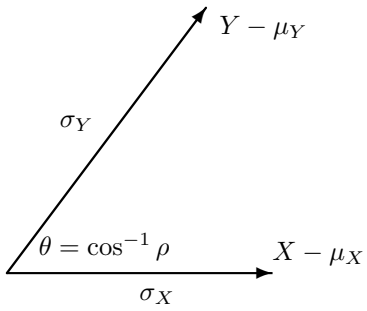

FIGURE 7.5 Random Variables as Vectors.

so the quantity

$$\theta = \cos^{-1}\Big(\frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}\Big) = \cos^{-1}\rho \tag{7.67}$$

can be thought of as the angle between the vectors. Here ρ is the correlation coefficient of the two random variables, so evidently

$$\rho = \cos(\theta)\ . \tag{7.68}$$

Thus, the correlation coefficient is the cosine of the angle between the vectors. It is therefore not surprising at all that

$$-1 \leq \rho \leq 1\ . \tag{7.69}$$

When ρ is near 1, the vectors are nearly aligned in the same direction, whereas when ρ is near -1 they are close to being oppositely aligned. The correlation coefficient is zero when these vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ (which represent the centered random variables) are orthogonal, or equivalently, the corresponding random variables have zero covariance,

$$\sigma_{X,Y} = 0\ . \tag{7.70}$$

6.011 Introduction to Communication, Control, and Signal Processing
Spring 2010