

CHAPTER 2

Probability theory

The *most* we can know is in terms of probabilities.

—Richard Feynman [Fey63, p. 6-11]

While writing my book [*Stochastic Processes*, first published in 1953] I had an argument with Feller. He asserted that everyone said “random variable” and I asserted that everyone said “chance variable.” We obviously had to use the same name in our books, so we decided the issue by a stochastic procedure. That is, we tossed for it and he won.

—Joseph Doob [Sne97, p. 307]

Probabilities do not exist.

—Bruno de Finetti [deF74]

In our attempt to filter a signal, we will be trying to extract meaningful information from a noisy signal. In order to accomplish this, we need to know something about what noise is, some of its characteristics, and how it works. This chapter reviews probability theory. We begin by discussing the basic concept of probability in Section 2.1, and then move on to random variables (RVs) in Section 2.2. The chapter then continues with the following topics:

- An RV is a general case of the normal scalars that we are familiar with, and so just as we can apply a functional mapping to a number, we can also apply

a functional mapping to an RV. We discuss functions (transformations) of random variables in Section 2.3.

- Just as we can have vectors of numbers, we can also have vectors of RVs, and so we discuss groups of random variables and random vectors in Section 2.4.
- Just as we can have scalar functions of time, we can also have RVs that are functions of time, and so we discuss RVs that change with time (stochastic processes) in Section 2.5.
- Stochastic processes can be divided into two categories: white noise and colored noise, and we discuss these concepts in Section 2.6.

We conclude in Section 2.7 with a high-level discussion of how to write a computer simulation of a noise process.

This chapter is only a brief introduction and review of probability and stochastic processes, and more detail can be found in many other books on the subject, such as [Pap02, Pee01].

2.1 PROBABILITY

How shall we define the concept of probability? Suppose we run an experiment a certain number of times. Sometimes event A occurs and sometimes it does not. For instance, our experiment may be rolling a six-sided die. Event A may be defined as the number 4 showing up on the top surface of the die after we roll the die. Common sense tells us that the probability of event A occurring is $1/6$. Likewise, we would expect that if we run our experiment many times, then we would see the number 1 appearing about $1/6$ of the time. This intuitive explanation forms the basis for our formal description of the concept of probability. We define the probability of event A as

$$P(A) = \frac{\text{Number of times } A \text{ occurs}}{\text{Total number of outcomes}} \quad (2.1)$$

This commonsense understanding of probability is called the relative frequency definition. A more formal and mathematically rigorous definition of probability can be obtained using set theory [Bil95, Nel87], which was pioneered by Andrey Kolomogorov in the 1930s. But for our purposes, the relative frequency definition is adequate.

In general, we know that there are n -choose- k different ways of selecting k objects from a total of n objects (assuming that the order of the objects does not matter), where n -choose- k is denoted and defined as

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \quad (2.2)$$

For instance, suppose we have a penny (P), nickel (N), dime (D), and quarter (Q). How many distinct subsets of three coins can we pick from that set? We can pick PND, PNQ, PDQ, or NDQ, for a total of four possible subsets. This is equal to 4-choose-3.

■ EXAMPLE 2.1

What is the probability of being dealt four of a kind¹ in poker? The total number of possible poker hands can be computed as the total number of subsets of size five that can be picked from a deck of 52 cards. The total number of possible hands is 52-choose-5 = 2,598,960. Out of all those hands, there are 48 possible hands containing four aces, 48 possible hands containing four kings, and so on. So there are a total of 13×48 hands containing four of a kind. Therefore the probability of being dealt four of a kind is

$$\begin{aligned} P(A) &= \frac{(13)(48)}{2,598,960} \\ &= 1/4165 \\ &\approx 0.024\% \end{aligned} \tag{2.3}$$

▽▽▽

The conditional probability of event A given event B can be defined if the probability of B is nonzero. The conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{2.4}$$

$P(A|B)$ is the conditional probability of A given B , that is, the probability that A occurs given the fact that B occurred. $P(A, B)$ is the joint probability of A and B , that is, the probability that events A and B both occur. The probability of a single event [for instance, $P(A)$ or $P(B)$] is called an *a priori* probability because it applies to the probability of an event apart from any previously known information. A conditional probability [for instance, $P(A|B)$] is called an *a posteriori* probability because it applies to a probability given the fact that some information about a possibly related event is already known.

For example, suppose that A is the appearance of a 4 on a die, and B is the appearance of an even number on a die. $P(A) = 1/6$. But if we know that the die has an even number on it, then $P(A) = 1/3$ (since the even number could be either a 2, 4, or 6). This example is intuitive, but we can also obtain the answer using Equation (2.4). $P(A, B)$ is the probability that both A occurs (we roll a 4) and B occurs (we roll an even number), so $P(A, B) = 1/6$. So Equation (2.4) gives

$$\begin{aligned} P(A|B) &= \frac{1/6}{1/2} \\ &= 1/3 \end{aligned} \tag{2.5}$$

The *a priori* probability of A is $1/6$. But the *a posteriori* probability of A given B is $1/3$.

■ EXAMPLE 2.2

Consider the eight shapes in Figure 2.1. We have three circles and five squares, so $P(\text{circle}) = 3/8$. Only one of the shapes is a gray circle, so $P(\text{gray, circle})$

¹Once I was dealt four sevens while playing poker with some friends (unfortunately, I was not playing for money at the time). I don't expect to see it again in my lifetime.

$= 1/8$. Of the three circles, only one is gray, so $P(\text{gray} | \text{circle}) = 1/3$. This last probability can be computed using Equation (2.4) as

$$\begin{aligned} P(\text{gray}|\text{circle}) &= \frac{P(\text{gray, circle})}{P(\text{circle})} \\ &= \frac{1/8}{3/8} \\ &= 1/3 \end{aligned} \tag{2.6}$$

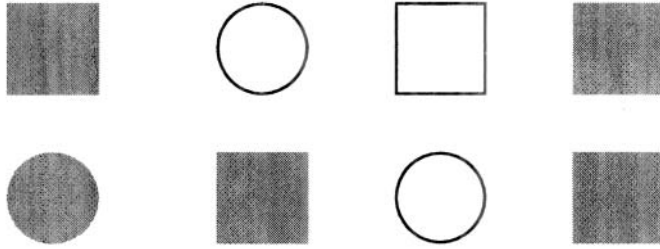


Figure 2.1 Some shapes for illustrating probability and Bayes' Rule.

▽▽▽

Note that we can use Equation (2.4) to write $P(B|A) = P(A, B)/P(A)$. We can solve both this equation and Equation (2.4) for $P(A, B)$ and equate the two expressions for $P(A, B)$ to obtain Bayes' Rule.

$$P(A|B)P(B) = P(B|A)P(A) \tag{2.7}$$

Bayes' Rule is often written by rearranging the above equation to obtain

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.8}$$

As an example, consider Figure 2.1. The probability of picking a gray shape given the fact that the shape is a circle can be computed from Bayes' Rule as

$$\begin{aligned} P(\text{gray}|\text{circle}) &= \frac{P(\text{circle}|\text{gray})P(\text{gray})}{P(\text{circle})} \\ &= \frac{(1/5)(5/8)}{3/8} \\ &= 1/3 \end{aligned} \tag{2.9}$$

We say that two events are independent if the occurrence of one event has no effect on the probability of the occurrence of the other event. For example, if A is the appearance of a 4 after rolling a die, and B is the appearance of a 3 after rolling *another* die, then A and B are independent. Mathematically, independence of A and B can be expressed several different ways. For example, we can write

$$\begin{aligned} P(A, B) &= P(A)P(B) \\ P(A|B) &= P(A) \\ P(B|A) &= P(B) \end{aligned} \tag{2.10}$$

if A and B are independent. As an example, recall from Equation (2.5) that if A is the appearance of a 4 on a die, and B is the appearance of an even number on a die, then $P(A) = 1/6$ and $P(A|B) = 1/3$. Since $P(A|B) \neq P(A)$ we see that A and B are dependent events.

2.2 RANDOM VARIABLES

We define a random variable (RV) as a functional mapping from a set of experimental outcomes (the domain) to a set of real numbers (the range). For example, the roll of a die can be viewed as a RV if we map the appearance of one dot on the die to the output one, the appearance of two dots on the die to the output two, and so on.

Of course, *after* we throw the die, the value of the die is no longer a random variable – it becomes certain. The outcome of a particular experiment is not an RV. If we define X as an RV that represents the roll of a die, then the probability that X will be a four is equal to $1/6$. If we then roll a four, the four is a realization of the RV X . If we then roll the die again and get a three, the three is another realization of the RV X . However, the RV X exists independently of any of its realizations. This distinction between an RV and its realizations is important for understanding the concept of probability. Realizations of an RV are not equal to the RV itself. When we say that the probability of $X = 4$ is equal to $1/6$, that means that there is a 1 out of 6 chance that each realization of X will be equal to 4. However, the RV X will always be random and will never be equal to a specific value.

An RV can be either continuous or discrete. The throw of a die is a discrete random variable because its realizations belong to a discrete set of values. The high temperature tomorrow is a continuous random variable because its realizations belong to a continuous set of values.

The most fundamental property of an RV X is its probability distribution function (PDF) $F_X(x)$, defined as

$$F_X(x) = P(X \leq x) \quad (2.11)$$

In the above equation, $F_X(x)$ is the PDF of the RV X , and x is a nonrandom independent variable or constant. Some properties of the PDF that can be obtained from its definition are

$$\begin{aligned} F_X(x) &\in [0, 1] \\ F_X(-\infty) &= 0 \\ F_X(\infty) &= 1 \\ F_X(a) &\leq F_X(b) \quad \text{if } a \leq b \\ P(a < X \leq b) &= F_X(b) - F_X(a) \end{aligned} \quad (2.12)$$

The probability density function (pdf) $f_X(x)$ is defined as the derivative of the PDF.

$$f_X(x) = \frac{dF_X(x)}{dx} \quad (2.13)$$

Some properties of the pdf that can be obtained from this definition are

$$\begin{aligned}
 F_X(x) &= \int_{-\infty}^x f_X(z) dz \\
 f_X(x) &\geq 0 \\
 \int_{-\infty}^{\infty} f_X(x) dx &= 1 \\
 P(a < x \leq b) &= \int_a^b f_X(x) dx
 \end{aligned} \tag{2.14}$$

The Q -function of an RV is defined as one minus the PDF. This is equal to the probability that the RV is greater than the argument of the function:

$$\begin{aligned}
 Q(x) &= 1 - F_X(x) \\
 &= P(X > x)
 \end{aligned} \tag{2.15}$$

Just as we spoke about conditional probabilities in Equation (2.4), we can also speak about the conditional PDF and the conditional pdf. The conditional distribution and density of the RV X given the fact that event A occurred are defined as

$$\begin{aligned}
 F_X(x|A) &= P(X \leq x|A) \\
 &= \frac{P(X \leq x, A)}{P(A)} \\
 f_X(x|A) &= \frac{dF_X(x|A)}{dx}
 \end{aligned} \tag{2.16}$$

Bayes' Rule, discussed in Section 2.1, can be generalized to conditional densities. Suppose we have random variables X_1 and X_2 . The conditional pdf of the RV X_1 given the fact that RV X_2 is equal to the realization x_2 is defined as

$$\begin{aligned}
 f_{X_1|X_2}(x_1|x_2) &= P[(X_1 \leq x_1)|(X_2 = x_2)] \\
 &= \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)}
 \end{aligned} \tag{2.17}$$

Although this is not entirely intuitive, it can be derived without too much difficulty [Pap02, Pee01]. Now consider the following product of two conditional pdf's:

$$\begin{aligned}
 f[x_1|(x_2, x_3, x_4)]f[(x_2, x_3)|x_4] &= \frac{f(x_1, x_2, x_3, x_4)}{f(x_2, x_3, x_4)} \frac{f(x_2, x_3, x_4)}{f(x_4)} \\
 &= \frac{f(x_1, x_2, x_3, x_4)}{f(x_4)} \\
 &= f[(x_1, x_2, x_3)|x_4]
 \end{aligned} \tag{2.18}$$

Note that in the above equation we have dropped the subscripts on the $f(\cdot)$ functions for ease of notation. This is commonly done if the random variable associated with the pdf is clear from the context. This is called the Chapman–Kolmogorov equation [Pap02]. It can be extended to any number of RVs and is fundamental to the Bayesian approach to state estimation (Chapter 15).

The expected value of an RV X is defined as its average value over a large number of experiments. This can also be called the expectation, the mean, or the average of

the RV. Suppose we run the experiment N times and observe a total of m different outcomes. We observe that outcome A_1 occurs n_1 times, A_2 occurs n_2 times, \dots , and A_m occurs n_m times. Then the expected value of X is computed as

$$E(X) = \frac{1}{N} \sum_{i=1}^m A_i n_i \quad (2.19)$$

$E(X)$ is also often written as $E(x)$, \bar{X} , or \bar{x} .

At this point, we will begin to use lowercase x instead of uppercase X when the meaning is clear. We have been using uppercase X to refer to an RV, and lowercase x to refer to a realization of the RV, which is a constant or independent variable. However, it should be clear that, for example, $E(x)$ is the expected value of the RV X , and so we will interchange x and X in order to simplify notation.

As an example of the expected value of an RV, suppose that we roll a die an infinite number of times. We would expect to see each possible number (one through six) $1/6$ of the time each. We can compute the expected value of the roll of the die as

$$\begin{aligned} E(X) &= \lim_{N \rightarrow \infty} \frac{1}{N} [(1)(N/6) + \dots + (6)(N/6)] \\ &= 3.5 \end{aligned} \quad (2.20)$$

Note that the expected value of an RV is not necessarily what we would expect to see when we run a particular experiment. For example, even though the above expected value of X is 3.5, we will never see a 3.5 when we roll a die.

We can also talk about a function of an RV, just as we can talk about a function of any scalar. (We will discuss this in more detail in Section 2.3.) If a function, say $g(X)$, acts upon an RV, then the output of the function is also an RV. For example, if X is the roll of a die, then $P(X = 4) = 1/6$. If $g(X) = X^2$, then $P[g(X) = 16] = 1/6$. We can compute the expected value of any function $g(X)$ as

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad (2.21)$$

where $f_X(x)$ is the pdf of X . If $g(X) = X$, then we can compute the expected value of X as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx \quad (2.22)$$

The variance of an RV is a measure of how much we expect the RV to vary from its mean. The variance is a measure of how much variability there is in an RV. In the extreme case, if the RV X always is equal to one value (for example, the die is loaded and we always get a 4 when we roll the die), then the variance of X is equal to 0. On the other extreme, if X can take on any value between $\pm\infty$ with equal probability, then the variance of X is equal to ∞ . The variance of an RV is formally defined as

$$\begin{aligned} \sigma_X^2 &= E[(X - \bar{x})^2] \\ &= \int_{-\infty}^{\infty} (x - \bar{x})^2 f_X(x) dx \end{aligned} \quad (2.23)$$

The standard deviation of an RV is σ , which is the square root of the variance. Sometimes we denote the standard deviation as σ_X if we need to be explicit about the RV whose standard deviation we are discussing. Note that the variance can be written as

$$\begin{aligned}\sigma^2 &= E[X^2 - 2X\bar{x} + \bar{x}^2] \\ &= E(X^2) - 2\bar{x}^2 + \bar{x}^2 \\ &= E(X^2) - \bar{x}^2\end{aligned}\quad (2.24)$$

We use the notation

$$X \sim (\bar{x}, \sigma^2) \quad (2.25)$$

to indicate that X is an RV with a mean of \bar{x} and a variance of σ^2 .

The skew of an RV is a measure of the asymmetry of the pdf around its mean. Skew is defined as

$$\text{skew} = E[(X - \bar{x})^3] \quad (2.26)$$

The skewness, also called the coefficient of skewness, is the skew normalized by the cube of the standard deviation:

$$\text{skewness} = \text{skew}/\sigma^3 \quad (2.27)$$

In general, the i th moment of a random variable X is the expected value of the i th power of X . The i th central moment of a random variable X is the expected value of the i th power of X minus its mean:

$$\begin{aligned}i\text{th moment of } X &= E(X^i) \\ i\text{th central moment of } X &= E[(X - \bar{x})^i]\end{aligned}\quad (2.28)$$

For example, the first moment of a random variable is equal to its mean. The first central moment of a random variable is always equal to 0. The second central moment of a random variable is equal to its variance.

An RV is called uniform if its pdf is a constant value between two limits. This indicates that the RV has an equally likely probability of obtaining any value between its limits, but a zero probability of obtaining a value outside of its limits:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \quad (2.29)$$

Figure 2.2 shows the pdf of an RV that is uniformly distributed between ± 1 . Note that the area of this curve is one (as is the area of all pdf's).

■ EXAMPLE 2.3

In this example we will find the mean and variance of an RV that is uniformly distributed between 1 and 3. The pdf of the RV is given as

$$f_X(x) = \begin{cases} 1/2 & x \in [1, 3] \\ 0 & \text{otherwise} \end{cases} \quad (2.30)$$

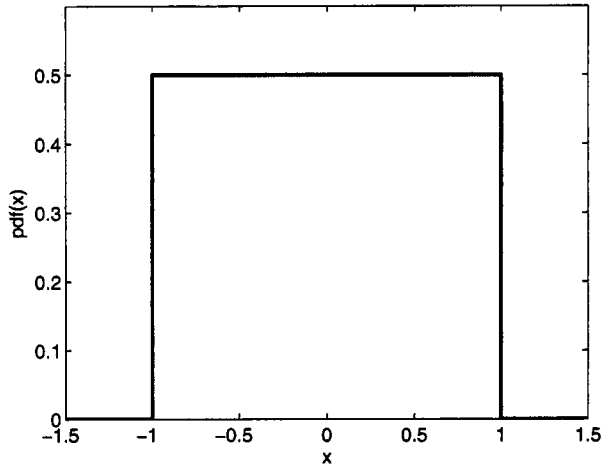


Figure 2.2 Probability density function of an RV uniformly distributed between ± 1 .

The mean is computed as follows:

$$\begin{aligned}
 \bar{x} &= \int_{-\infty}^{\infty} x f_X(x) dx \\
 &= \int_{-1}^1 \frac{1}{2} x dx \\
 &= 0
 \end{aligned} \tag{2.31}$$

The variance is computed as follows:

$$\begin{aligned}
 \sigma_X^2 &= \int_{-\infty}^{\infty} \frac{1}{2} (x - \bar{x})^2 f(x) dx \\
 &= \int_{-1}^1 \frac{1}{2} (x - 0)^2 dx \\
 &= \frac{1}{3}
 \end{aligned} \tag{2.32}$$

▽▽▽

An RV is called Gaussian or normal if its pdf is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[\frac{-(x - \bar{x})^2}{2\sigma^2} \right] \tag{2.33}$$

This is called the Laplace distribution in France, but it had many other discoverers, including Robert Adrain. Note that \bar{x} and σ in the above pdf are the mean and standard deviation of the Gaussian RV. We use the notation

$$X \sim N(\bar{x}, \sigma^2) \tag{2.34}$$

to indicate that X is a Gaussian RV with a mean of \bar{x} and a variance of σ^2 . Figure 2.3 shows the pdf of a Gaussian RV with a mean of zero and a variance

of one. If the mean changes, the pdf will shift to the left or right. If the variance increases, the pdf will spread out. If the variance decreases, the pdf will be squeezed in. The PDF of a Gaussian RV is given by

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp[-(z - \bar{x})^2/2\sigma^2] dz \quad (2.35)$$

This integral does not have a closed-form solution, and so it must be evaluated numerically. However, its evaluation can be simplified by considering the normalized Gaussian PDF of an RV with zero mean and unity variance:

$$F_{X0}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-z^2/2) dz \quad (2.36)$$

It can be shown that any Gaussian PDF can be expressed in terms of this normalized PDF as

$$F_X(x) = F_{X0}\left(\frac{x - \bar{x}}{\sigma}\right) \quad (2.37)$$

In addition, a Gaussian PDF can be approximated as the following closed-form expression [Bor79]:

$$\begin{aligned} F_X(x) &\approx 1 - \left[\frac{1}{(1-a)x + a\sqrt{x^2+b}} \right] \frac{\exp(-x^2/2)}{\sqrt{2\pi}} & x \geq 0 \\ a &= 0.339 \\ b &= 5.510 \end{aligned} \quad (2.38)$$

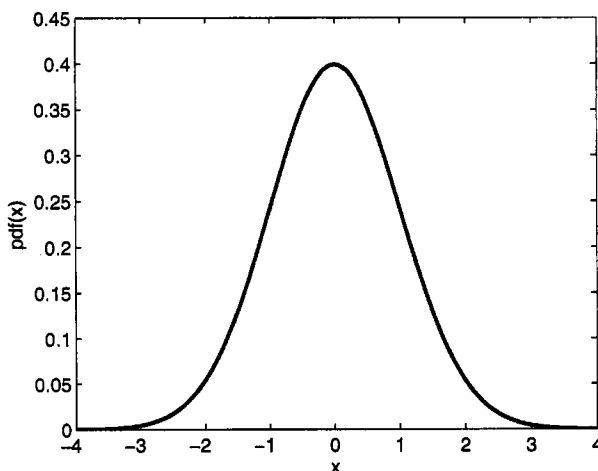


Figure 2.3 Probability density function of a Gaussian RV with a mean of zero and a variance of one.

Suppose we have a random variable X with a mean of zero and a symmetric pdf [i.e., $f_X(x) = f_X(-x)$]. This is the case, for example, for the pdf's shown in

Figures 2.2 and 2.3. In this case, the i th moment of X can be written as

$$\begin{aligned}
 m_i &= E(X^i) \\
 &= \int_{-\infty}^{\infty} x^i f_X(x) dx \\
 &= \int_{-\infty}^0 x^i f_X(x) dx + \int_0^{\infty} x^i f_X(x) dx
 \end{aligned} \tag{2.39}$$

If i is odd then $x^i = -(-x)^i$. Combined with the fact that $f_X(x) = f_X(-x)$, we see that

$$\begin{aligned}
 \int_{-\infty}^0 x^i f_X(x) dx &= \int_0^{\infty} (-x)^i f_X(-x) dx \\
 &= - \int_0^{\infty} x^i f_X(x) dx
 \end{aligned} \tag{2.40}$$

So for odd i , the i th moment in Equation (2.39) is zero. We see that all of the odd moments of a zero-mean random variable with a symmetric pdf are equal to 0.

2.3 TRANSFORMATIONS OF RANDOM VARIABLES

In this section, we will look at what happens to the pdf of an RV when we pass the RV through some function. Suppose that we have two RVs, X and Y , related to one another by the monotonic² functions $g(\cdot)$ and $h(\cdot)$:

$$\begin{aligned}
 Y &= g(X) \\
 X &= g^{-1}(Y) = h(Y)
 \end{aligned} \tag{2.41}$$

If we know the pdf of X [$f_X(x)$], then we can compute the pdf of Y [$f_Y(y)$] as follows:

$$\begin{aligned}
 P(X \in [x, x+dx]) &= P(Y \in [y, y+dy]) \quad (dx > 0) \\
 \int_x^{x+dx} f_X(z) dz &= \begin{cases} \int_y^{y+dy} f_Y(z) dz & \text{if } dy > 0 \\ - \int_y^{y+dy} f_Y(z) dz & \text{if } dy < 0 \end{cases} \\
 f_X(x) dx &= f_Y(y) |dy| \\
 f_Y(y) &= \left| \frac{dx}{dy} \right| f_X[h(y)] \\
 &= |h'(y)| f_X[h(y)]
 \end{aligned} \tag{2.42}$$

where we have used the assumption of small dx and dy in the above calculation.

²A monotonic function is a function whose slope is either always nonnegative or always nonpositive. If the slope is always nonnegative, then the function is monotonically nondecreasing. If the slope is always positive, then the function is monotonically increasing. If the slope is always nonpositive, then the function is monotonically nonincreasing. If the slope is always negative, then the function is monotonically decreasing.

■ EXAMPLE 2.4

In this example, we will find the pdf of a linear function of a Gaussian RV. Suppose that $X \sim N(\bar{x}, \sigma_x^2)$ and $Y = g(X) = aX + b$, where $a \neq 0$ and b are any real constants. Then

$$\begin{aligned}
 X &= h(Y) \\
 &= (Y - b)/a \\
 h'(y) &= 1/a \\
 f_Y(y) &= |h'(y)| f_X[h(y)] \\
 &= \left| \frac{1}{a} \right| \frac{1}{\sigma_X \sqrt{2\pi}} \exp \left\{ \frac{-[(y - b)/a - \bar{x}]^2}{2\sigma_X^2} \right\} \\
 &= \frac{1}{a\sigma_X \sqrt{2\pi}} \exp \left\{ \frac{-[y - (a\bar{x} + b)]^2}{2a^2\sigma_X^2} \right\}
 \end{aligned} \tag{2.43}$$

In other words, the RV Y is Gaussian with a mean and variance given by

$$\begin{aligned}
 \bar{y} &= a\bar{x} + b \\
 \sigma_Y^2 &= a^2\sigma_X^2
 \end{aligned} \tag{2.44}$$

This important example shows that a linear transformation of a Gaussian RV results in a new Gaussian RV.

▽▽▽

■ EXAMPLE 2.5

Suppose that we pass a Gaussian RV $X \sim N(0, \sigma_x^2)$ through the nonlinear function $Y = g(X) = X^3$:

$$\begin{aligned}
 X &= h(Y) \\
 &= Y^{1/3} \\
 h'(y) &= \frac{y^{-2/3}}{3} \\
 f_Y(y) &= |h'(y)| f_X[h(y)] \\
 &= \frac{y^{-2/3}}{3} \frac{1}{\sigma_x \sqrt{2\pi}} \exp[-x^2/(2\sigma_x^2)] \\
 &= \frac{y^{-2/3}}{3} \frac{1}{\sigma_x \sqrt{2\pi}} \exp[-y^{2/3}/(2\sigma_x^2)]
 \end{aligned} \tag{2.45}$$

We see that the nonlinear transformation $Y = X^3$ converts a Gaussian RV to a non-Gaussian RV. It can be seen that $f_Y(y)$ approaches ∞ as $y \rightarrow 0$. Nevertheless, the area under the $f_Y(y)$ curve is equal to 1 since it is a pdf.

▽▽▽

In the more general case of RVs related by the function $Y = g(X)$, where $g(\cdot)$ is a nonmonotonic function, the pdf of Y (evaluated at y) can be computed from the pdf of X as

$$f_Y(y) = \sum_i f_X(x_i) / |g'(x_i)| \tag{2.46}$$

where the x_i values are the solutions of the equation $y = g(x)$.

2.4 MULTIPLE RANDOM VARIABLES

We have already defined the probability distribution function of an RV. For example, if X and Y are RVs, then their distribution functions are defined as

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ F_Y(y) &= P(Y \leq y) \end{aligned} \quad (2.47)$$

Now we define the probability that both $X \leq x$ and $Y \leq y$ as the joint probability distribution function of X and Y :

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \quad (2.48)$$

If the meaning is clear from the context, we often use the shorthand notation $F(x, y)$ to represent the distribution function $F_{XY}(x, y)$. Some properties of the joint distribution function are

$$\begin{aligned} F(x, y) &\in [0, 1] \\ F(x, -\infty) = F(-\infty, y) &= 0 \\ F(\infty, \infty) &= 1 \\ F(a, c) &\leq F(b, d) \quad \text{if } a \leq b \text{ and } c \leq d \\ P(a < x \leq b, c < y \leq d) &= F(b, d) + F(a, c) - F(a, d) - F(b, c) \\ F(x, \infty) &= F(x) \\ F(\infty, y) &= F(y) \end{aligned} \quad (2.49)$$

Note from the last two properties that the distribution function of one RV can be obtained from the joint distribution function. When the distribution function for a single RV is obtained this way it is called the marginal distribution function.

The joint probability density function is defined as the following derivative of the joint PDF:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad (2.50)$$

As before, we often use the shorthand notation $f(x, y)$ to represent the density function $f_{XY}(x, y)$. Some properties of the joint pdf that can be obtained from this definition are

$$\begin{aligned} F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(z_1, z_2) dz_1 dz_2 \\ f(x, y) &\geq 0 \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= 1 \\ P(a < x \leq b, c < y \leq d) &= \int_c^d \int_a^b f(x, y) dx dy \\ f(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ f(y) &= \int_{-\infty}^{\infty} f(x, y) dx \end{aligned} \quad (2.51)$$

Note from the last two properties that the density function of one RV can be obtained from the joint density function. When the density function for a single RV is obtained this way it is called the marginal density function. Computing the expected value of a function $g(\cdot, \cdot)$ of two RVs is similar to computing the expected value of a function of a single RV:

$$E[g(x, y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \quad (2.52)$$

2.4.1 Statistical independence

Recall from Section 2.1 that two events are independent if the occurrence of one event has no effect on the probability of the occurrence of the other event. We extend this to say that RVs X and Y are independent if they satisfy the following relation:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \quad \text{for all } x, y \quad (2.53)$$

From our definition of joint distribution and density functions, we see that this implies

$$\begin{aligned} F_{XY}(x, y) &= F_X(x)F_Y(y) \\ f_{XY}(x, y) &= f_X(x)f_Y(y) \end{aligned} \quad (2.54)$$

The central limit theorem says that the sum of independent RVs tends toward a Gaussian RV, regardless of the pdf of the individual RVs that contribute to the sum. This is why so many RVs in nature seem to have a Gaussian distribution. Many RVs in nature are actually the sum of many individual and independent RVs. For example, the high temperature on any given day in any given location tends to follow a Gaussian distribution. This is because the high temperature is affected by clouds, precipitation, wind, air pressure, humidity, and other factors. Each of these factors is in turn determined by other random factors. The combination of many independent random variables determines the high temperature, which has a Gaussian pdf.

We define the covariance of two scalar RVs X and Y as

$$\begin{aligned} C_{XY} &= E[(X - \bar{X})(Y - \bar{Y})] \\ &= E(XY) - \bar{X}\bar{Y} \end{aligned} \quad (2.55)$$

We define the correlation coefficient of two scalar RVs X and Y as

$$\rho = \frac{C_{XY}}{\sigma_x \sigma_y} \quad (2.56)$$

The correlation coefficient is a normalized measurement of the independence between two RVs X and Y . If X and Y are independent, then $\rho = 0$ (although the converse is not necessarily true). If Y is a linear function of X then $\rho = \pm 1$ (see Problem 2.9).

We define the correlation of two scalar RVs X and Y as

$$R_{XY} = E(XY) \quad (2.57)$$

Two RVs are said to be uncorrelated if $R_{XY} = E(X)E(Y)$.

From the definition of independence, we see that if two RVs are independent then they are also uncorrelated. Independence implies uncorrelatedness, but uncorrelatedness does not necessarily imply independence. However, in the special case in which two RVs are both Gaussian and uncorrelated, then it follows that they are also independent.

Two RVs are said to be orthogonal if $R_{XY} = 0$. If two RVs are uncorrelated, then they are orthogonal only if at least one of them is zero-mean. If two RVs are orthogonal, then they may or may not be uncorrelated.

■ EXAMPLE 2.6

Two rolls of the dice are represented by the RVs X and Y . The two RVs are independent because one roll of the die does not have any effect on a second roll of the die. Each roll of the die has an equally likely probability ($1/6$) of being a 1, 2, 3, 4, 5, or 6. Therefore,

$$\begin{aligned} E(X) = E(Y) &= \frac{1 + 2 + 3 + 4 + 5 + 6}{6} \\ &= 3.5 \end{aligned} \quad (2.58)$$

There are 36 possible combinations of the two rolls of the die. We could get the combination (1,1), (1,2), and so on. Each of these 36 combinations have an equally likely probability ($1/36$). Therefore, the correlation between X and Y is

$$\begin{aligned} R_{XY} = E(XY) &= \frac{1}{36} \sum_{i=1}^6 \sum_{j=1}^6 ij \\ &= 12.25 \\ &= E(X)E(Y) \end{aligned} \quad (2.59)$$

Since $E(XY) = E(X)E(Y)$, we see that X and Y are uncorrelated. However, $R_{XY} \neq 0$, so X and Y are not orthogonal.

▽▽▽

■ EXAMPLE 2.7

A slot machine is rigged so you get 1 or -1 with equal probability the first spin X , and the opposite number the second spin Y . We have equal probabilities of obtaining (X, Y) outcomes of (1, -1) and (-1 , 1). The two RVs are dependent because the realization of Y depends on the realization of X . We also see that

$$\begin{aligned} E(X) &= 0 \\ E(Y) &= 0 \\ E(XY) &= \frac{(1)(-1) + (-1)(1)}{2} \\ &= -1 \end{aligned} \quad (2.60)$$

We see that X and Y are correlated because $E(XY) \neq E(X)E(Y)$. We also see that X and Y are not orthogonal because $E(XY) \neq 0$.

▽▽▽

■ EXAMPLE 2.8

A slot machine is rigged so you get -1 , 0 , or $+1$ with equal probability the first spin X . On the second spin Y you get 1 if $X = 0$, and 0 if $X \neq 0$. The two RVs are dependent because the realization of Y depends on the realization of X . We also see that

$$\begin{aligned}
 E(X) &= \frac{-1 + 0 + 1}{3} \\
 &= 0 \\
 E(Y) &= \frac{0 + 1 + 0}{3} \\
 &= 1/3 \\
 E(XY) &= \frac{(-1)(0) + (0)(1) + (1)(0)}{3} \\
 &= 0
 \end{aligned} \tag{2.61}$$

We see that X and Y are uncorrelated because $E(XY) = E(X)E(Y)$. We also see that X and Y are orthogonal because $E(XY) = 0$. This example illustrates the fact that uncorrelatedness does not necessarily imply independence.

▽▽▽

■ EXAMPLE 2.9

Suppose that x and y are independent RVs, and the RV z is computed as $z = g(x) + h(y)$. In this example, we will calculate the mean of z :

$$\begin{aligned}
 E(z) &= E[g(x) + h(y)] \\
 &= \int \int [g(x) + h(y)] f(x, y) dx dy \\
 &= \int \int g(x) f(x) f(y) dx dy + \int \int h(y) f(x) f(y) dx dy \\
 &= \int g(x) f(x) dx \int f(y) dy + \int h(y) f(y) dy \int f(x) dx \\
 &= E[g(x)](1) + E[h(y)](1) \\
 &= E[g(x)] + E[h(y)]
 \end{aligned} \tag{2.62}$$

As a special case of this example, we see that the mean of the sum of two independent RVs is equal to the sum of their means. That is,

$$E(x + y) = E(x) + E(y) \quad \text{if } x \text{ and } y \text{ are independent} \tag{2.63}$$

▽▽▽

■ EXAMPLE 2.10

Suppose we roll a die twice. What is the expected value of the sum of the two outcomes? We use X and Y to refer to the two rolls of the die, and we use

Z to refer to the sum of the two outcomes. Therefore, $Z = X + Y$. Since X and Y are independent, we have

$$\begin{aligned} E(Z) &= E(X) + E(Y) \\ &= 3.5 + 3.5 \\ &= 7 \end{aligned} \quad (2.64)$$

▽▽▽

■ EXAMPLE 2.11

Consider the circuit of Figure 2.4. The input voltage V is uniformly distributed on $[-1, 1]$. Voltage V has units of volts, and the two currents have units of amps.

$$\begin{aligned} I_1 &= \begin{cases} 0 & \text{if } V > 0 \\ V & \text{if } V \leq 0 \end{cases} \\ I_2 &= \begin{cases} V & \text{if } V \geq 0 \\ 0 & \text{if } V < 0 \end{cases} \end{aligned} \quad (2.65)$$

We see that I_1 is uniformly distributed on $[-1, 0]$ and I_2 is uniformly distributed on $[0, 1]$. The RVs V , I_1 , and I_2 have expected values

$$\begin{aligned} E(V) &= 0 \\ E(I_1) &= -1/2 \\ E(I_2) &= 1/2 \end{aligned} \quad (2.66)$$

The RVs I_1 and I_2 are not independent because they are related to each other; if $I_2 \neq 0$ then $I_1 = 0$, and if $I_1 \neq 0$ then $I_2 = 0$. Since either I_1 or I_2 is equal to 0 at every time instant, $I_1 I_2 = 0$ and $E(I_1 I_2) = 0$. Therefore I_1 and I_2 are orthogonal. Since $E(I_1)E(I_2) = -1/4$, we see that $E(I_1 I_2) \neq E(I_1)E(I_2)$, and I_1 and I_2 are correlated.

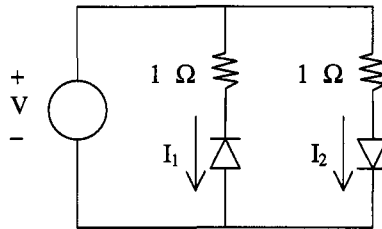


Figure 2.4 Circuit for Example 2.11.

▽▽▽

2.4.2 Multivariate statistics

The discussion in the previous subsection can be generalized for RVs that are vectors. In this case, the quantities defined earlier become vectors and matrices. Given

an n -element RV X and an m -element RV Y (assuming that both X and Y are column vectors), their correlation is defined as

$$\begin{aligned} R_{XY} &= E(XY^T) \\ &= \begin{bmatrix} E(X_1Y_1) & \cdots & E(X_1Y_m) \\ \vdots & & \vdots \\ E(X_nY_1) & \cdots & E(X_nY_m) \end{bmatrix} \end{aligned} \quad (2.67)$$

Their covariance is defined as

$$\begin{aligned} C_{XY} &= E[(X - \bar{X})(Y - \bar{Y})^T] \\ &= E(XY^T) - \bar{X}\bar{Y}^T \end{aligned} \quad (2.68)$$

The autocorrelation of the n -element RV X is defined as

$$\begin{aligned} R_X &= E[XX^T] \\ &= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} \end{aligned} \quad (2.69)$$

Note that $E(X_iX_j) = E(X_jX_i)$ so $R_X = R_X^T$. An autocorrelation matrix is always symmetric. Also note that for any n -element column vector z we have

$$\begin{aligned} z^T R_X z &= z^T E[XX^T] z \\ &= E[z^T XX^T z] \\ &= E[(z^T X)^2] \\ &\geq 0 \end{aligned} \quad (2.70)$$

So an autocorrelation matrix is always positive semidefinite.

The autocovariance of the n -element RV X is defined as

$$\begin{aligned} C_X &= E[(X - \bar{X})(X - \bar{X})^T] \\ &= \begin{bmatrix} E[(X_1 - \bar{X}_1)^2] & \cdots & E[(X_1 - \bar{X}_1)(X_n - \bar{X}_n)] \\ \vdots & & \vdots \\ E[(X_n - \bar{X}_n)(X_1 - \bar{X}_1)] & \cdots & E[(X_n - \bar{X}_n)^2] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{bmatrix} \end{aligned} \quad (2.71)$$

Note that $\sigma_{ij} = \sigma_{ji}$, so $C_X = C_X^T$. An autocovariance matrix is always symmetric. Also note that for any n -element column vector z we have

$$\begin{aligned} z^T C_X z &= z^T E[(X - \bar{X})(X - \bar{X})^T] z \\ &= E[z^T (X - \bar{X})(X - \bar{X})^T z] \\ &= E[(z^T (X - \bar{X}))^2] \\ &\geq 0 \end{aligned} \quad (2.72)$$

So an autocovariance matrix is always positive semidefinite.

An n -element RV X is Gaussian (normal)³ if

$$\text{pdf}(X) = \frac{1}{(2\pi)^{n/2}|C_X|^{1/2}} \exp \left[\frac{-1}{2}(X - \bar{X})^T C_X^{-1}(X - \bar{X}) \right] \quad (2.73)$$

Now consider a Gaussian RV X that undergoes a linear transformation:

$$\begin{aligned} Y &= g(X) \\ &= AX + b \end{aligned} \quad (2.74)$$

where A is a constant $n \times n$ matrix, and b is a constant n -element vector. If A is invertible, then

$$\begin{aligned} X &= h(Y) \\ &= A^{-1}Y - A^{-1}b \end{aligned} \quad (2.75)$$

From Equation (2.42) we obtain

$$\begin{aligned} f_Y(y) &= |h'(y)|f_X[h(y)] \\ &= |A^{-1}| \frac{1}{(2\pi)^{n/2}|C_X|^{1/2}} \exp \left[\frac{-1}{2}(A^{-1}y - A^{-1}b - \bar{x})^T C_X^{-1}(\dots) \right] \\ &= |A^{-1}| \frac{1}{(2\pi)^{n/2}|C_X|^{1/2}} \times \\ &\quad \exp \left[\frac{-1}{2}(A^{-1}y - A^{-1}b - A^{-1}A\bar{x})^T C_X^{-1}(\dots) \right] \\ &= \frac{1}{(2\pi)^{n/2}|A||C_X|^{1/2}} \times \\ &\quad \exp \left[\frac{-1}{2}(A^{-1}y - A^{-1}\bar{y})^T C_X^{-1}(A^{-1}y - A^{-1}\bar{y}) \right] \\ &= \frac{1}{(2\pi)^{n/2}|A|^{1/2}|C_X|^{1/2}|A^T|^{1/2}} \times \\ &\quad \exp \left[\frac{-1}{2}(y - \bar{y})^T A^{-T} C_X^{-1} A^{-1}(y - \bar{y}) \right] \\ &= \frac{1}{(2\pi)^{n/2}|AC_X A^T|^{1/2}} \exp \left[\frac{-1}{2}(y - \bar{y})^T (AC_X A^T)^{-1}(y - \bar{y}) \right] \\ y &\sim N(A\bar{x} + b, AC_X A^T) \end{aligned} \quad (2.76)$$

This shows that normality is preserved in linear transformations of random vectors (just as it is preserved in linear transformations of random scalars, as seen in Example 2.4).

³Francis Edgeworth (1845-1926), an Irish economist and mathematician, first provided a general description and study of the multivariate Gaussian probability distribution in 1892 [Sor80].

2.5 STOCHASTIC PROCESSES

A stochastic process, also called a random process, is a very simple generalization of the concept of an RV. A stochastic process $X(t)$ is an RV X that changes with time.⁴ A stochastic process can be one of four types.

- If the RV at each time is continuous and time is continuous, then $X(t)$ is a continuous random process. For example, the temperature at each moment of the day is a continuous random process because both temperature and time are continuous.
- If the RV at each time is discrete and time is continuous, then $X(t)$ is a discrete random process. For example, the number of people in a given building at each moment of the day is a discrete random process because the number of people is a discrete variable and time is continuous.
- If the RV at each time is continuous and time is discrete, then $X(t)$ is a continuous random sequence. For example, the high temperature each day is a continuous random sequence because temperature is continuous but time is discrete (day one, day two, etc.).
- If the RV at each time is discrete and time is discrete, then $X(t)$ is a discrete random sequence. For example, the highest number of people in a given building each day is a discrete random sequence because the number of people is a discrete variable and time is also discrete.

Since a stochastic process is an RV that changes with time, it has a distribution and density function that are functions of time. The PDF of $X(t)$ is

$$F_X(x, t) = P(X(t) \leq x) \quad (2.77)$$

If $X(t)$ is a random vector, then the inequality above is an element-by-element inequality. For example, if $X(t)$ has n elements, then

$$F_X(x, t) = P[X_1(t) \leq x_1 \text{ and } \cdots X_n(t) \leq x_n(t)] \quad (2.78)$$

The pdf of $X(t)$ is

$$f_X(x, t) = \frac{dF_X(x, t)}{dx} \quad (2.79)$$

If $X(t)$ is a random vector, then the derivative above is taken once with respect to each element of x . For example, if $X(t)$ has n elements, then

$$f_X(x, t) = \frac{d^n F_X(x, t)}{dx_1 \cdots dx_n} \quad (2.80)$$

The mean and covariance of $X(t)$ are also functions of time:

$$\bar{x}(t) = \int_{-\infty}^{\infty} x f(x, t) dx$$

⁴Actually, the independent variable does not have to be time; for example, it could be spatial location or something else. But typically the independent variable is time, and in this book it will always be time.

$$\begin{aligned}
 C_X(t) &= E \left\{ [X(t) - \bar{x}(t)] [X(t) - \bar{x}(t)]^T \right\} \\
 &= \int_{-\infty}^{\infty} [x - \bar{x}(t)] [x - \bar{x}(t)]^T f(x, t) dx
 \end{aligned} \tag{2.81}$$

Note that $X(t)$ at two different times (t_1 and t_2) comprise two different random variables $[X(t_1)$ and $X(t_2)]$. Therefore, we can talk about the joint distribution and joint density functions of $X(t_1)$ and $X(t_2)$. These are called the second-order distribution function and the second-order density function:

$$\begin{aligned}
 F(x_1, x_2, t_1, t_2) &= P(X(t_1) \leq x_1, X(t_2) \leq x_2) \\
 f(x_1, x_2, t_1, t_2) &= \frac{\partial^2 F(x_1, x_2, t_1, t_2)}{\partial x_1 \partial x_2}
 \end{aligned} \tag{2.82}$$

As discussed earlier, if $X(t)$ is an n -element random vector, then the inequality that defines $F(x_1, x_2, t_1, t_2)$ actually consists of $2n$ inequalities, and the derivative that defines $f(x_1, x_2, t_1, t_2)$ actually consists of $2n$ derivatives.

The correlation between the two RVs $X(t_1)$ and $X(t_2)$ is called the autocorrelation of the stochastic process $X(t)$:

$$R_X(t_1, t_2) = E [X(t_1)X^T(t_2)] \tag{2.83}$$

The autocovariance of a stochastic process is defined as

$$C_X(t_1, t_2) = E \left\{ [X(t_1) - \bar{X}(t_1)] [X(t_2) - \bar{X}(t_2)]^T \right\} \tag{2.84}$$

For some stochastic processes, the pdf does not change with time. For example, if we flip a coin ten times then we can view that process as a stochastic process with the statistics of the process being the same at each of the ten time instances. In this case, the stochastic process is called strict-sense stationary (SSS), or just stationary for short. In this case, the mean of the stochastic process is constant with respect to time, and the autocorrelation is a function of the time difference $t_2 - t_1$ (not a function of the absolute times):

$$\begin{aligned}
 E[X(t)] &= \bar{x} \\
 E[X(t_1)X^T(t_2)] &= R_X(t_2 - t_1)
 \end{aligned} \tag{2.85}$$

For some stochastic processes, these two conditions are true even though the pdf does change with time. Stochastic processes for which these two conditions are true are called wide-sense stationary (WSS). A stationary process is wide-sense stationary, but a wide-sense stationary process may or may not be stationary. From the definition of autocorrelation, it can be shown that for a wide-sense stationary process the following properties hold:

$$\begin{aligned}
 R_X(0) &= E[X(t)X^T(t)] \\
 R_X(-\tau) &= R_X(\tau)
 \end{aligned} \tag{2.86}$$

For scalar stochastic processes, it can be shown that

$$|R_X(\tau)| \leq R_X(0) \tag{2.87}$$

■ EXAMPLE 2.12

1. The high temperature each day can be considered a stochastic process. However, this process is not stationary. The high temperature on a day in July might be an RV with a mean of 100 degrees Fahrenheit, but the high temperature on a day in December might have a mean of 30 degrees. This is a stochastic process whose statistics change with time, so the process is not stationary.
2. Electrical noise in a voltmeter might have a mean of zero and a variance of one millivolt. If we come back the next day and measure the noise again, the mean and variance may be the same as before. If the statistics of the noise are the same every day, then the electrical noise is a stationary process. Note that in reality the noise statistics will eventually change. For example, after a few decades the instrument will begin degrading and the electrical noise mean and variance will change. In this sense, there is no such thing as a stationary random process. Eventually, the universe will freeze and all signals will change. But for practical purposes, if the statistics of a random process do not change over the time interval of interest, then we consider the process to be stationary.
3. Tomorrow's closing price of the Dow Jones Industrial Average might be an RV with a certain mean and variance. However, 100 years ago the closing price had a mean that was much lower. The closing price of the stock market is an RV whose mean generally increases with time. Therefore, the stock market price is a nonstationary stochastic process.

▽▽▽

Suppose we have a stochastic process $X(t)$. Further suppose that the process has a realization $x(t)$. The time average of $X(t)$ is denoted as $A[X(t)]$, and the time autocorrelation of $X(t)$ is denoted as $R[X(t)]$. These quantities are defined for continuous-time random processes as

$$\begin{aligned} A[X(t)] &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t) dt \\ R[X(t), \tau] &= A[X(t)X^T(t + \tau)] \end{aligned} \quad (2.88)$$

The definitions for discrete-time random processes are straightforward extensions of the continuous-time definitions.

An ergodic process is a stationary random process for which

$$\begin{aligned} A[X(t)] &= E(X) \\ R[X(t), \tau] &= R_X(\tau) \end{aligned} \quad (2.89)$$

In the real world, we are often limited to only a few realizations of a stochastic process. For example, if we measure the fluctuation of a voltmeter reading, we are actually only measuring only one realization of a stochastic process. We can compute the time average, time autocorrelation, and other time-based statistics of the realization. If the random process is ergodic, then we can use those time averages to estimate the statistics of the stochastic process.

■ EXAMPLE 2.13

1. Suppose each unit of an electrical instrument is manufactured with a small random bias. Is the noise of the instrumentation ergodic? If we measure the noise of one instrument then we measure its bias, which is equal to its mean. However, if we measure the noise of another instrument it might have a different mean because it has a different bias. In other words, we cannot obtain the mean of the stochastic process by simply investigating one instrument (i.e., one realization of the stochastic process). Therefore, the stochastic process is not ergodic.
2. Suppose each unit of an electrical instrument is manufactured identically, each with zero-mean stationary Gaussian noise. Is the noise ergodic? In this case we could measure the mean of the process by measuring the noise of many separate instruments at one instant of time, or by measuring the noise of one instrument over an extended period of time. Either experiment would correctly inform us that the mean of the stochastic process is zero. We could find the statistics of the stochastic process using all the instruments at a single time, or using a single instrument at many different times. Therefore, the stochastic process is ergodic.

▽▽▽

The definitions of correlation and covariance can be extended to two stochastic processes $X(t)$ and $Y(t)$. The cross correlation of $X(t)$ and $Y(t)$ is defined as

$$R_{XY}(t_1, t_2) = E[X(t_1)Y^T(t_2)] \quad (2.90)$$

Two random processes $X(t)$ and $Y(t)$ are said to be uncorrelated if $R_{XY}(t_1, t_2) = E[X(t_1)]E[Y^T(t_2)]$ for all t_1 and t_2 . The cross covariance of $X(t)$ and $Y(t)$ is defined as

$$C_{XY}(t_1, t_2) = E\{[X(t_1) - \bar{X}(t_1)][Y(t_2) - \bar{Y}(t_2)]^T\} \quad (2.91)$$

2.6 WHITE NOISE AND COLORED NOISE

If the RV $X(t_1)$ is independent from the RV $X(t_2)$ for all $t_1 \neq t_2$ then $X(t)$ is called white noise. Otherwise, $X(t)$ is called colored noise.

The whiteness or color content of a stochastic process can be characterized by its power spectrum. The power spectrum $S_X(\omega)$ of a wide-sense stationary stochastic process $X(t)$ is defined as the Fourier transform of the autocorrelation. The autocorrelation is the inverse Fourier transform of the power spectrum.

$$\begin{aligned} S_X(\omega) &= \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau \\ R_X(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) e^{j\omega\tau} d\omega \end{aligned} \quad (2.92)$$

These equations are called the Wiener–Khinchine relations after Norbert Wiener and Aleksandr Khinchin. Note that some authors put the term $1/2\pi$ on the right side of the $S_X(\omega)$ definition, in which case the $1/2\pi$ term on the right side of the

$R_X(\tau)$ definition disappears. The power spectrum is sometimes referred to as the power density spectrum, the power spectral density, or the power density. The power of a wide-sense stationary stochastic process is defined as

$$P_X = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) d\omega \quad (2.93)$$

The cross power spectrum of two wide-sense stationary stochastic processes $X(t)$ and $Y(t)$ is the Fourier transform of the cross correlation:

$$\begin{aligned} S_{XY}(\omega) &= \int_{-\infty}^{\infty} R_{XY}(\tau) e^{-j\omega\tau} d\tau \\ R_{XY}(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{XY}(\omega) e^{j\omega\tau} d\omega \end{aligned} \quad (2.94)$$

Similar definitions hold for discrete-time random processes. The power spectrum of a discrete-time random process is defined as

$$\begin{aligned} S_X(\omega) &= \sum_{k=-\infty}^{\infty} R_X(k) e^{-j\omega k} \quad \omega \in [-\pi, \pi] \\ R_X(k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_X(\omega) e^{jk\omega} d\omega \end{aligned} \quad (2.95)$$

A discrete-time stochastic process $X(t)$ is called white noise if

$$\begin{aligned} R_X(k) &= \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \\ &= \sigma^2 \delta_k \end{aligned} \quad (2.96)$$

where δ_k is the Kronecker delta function, defined as

$$\delta_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \quad (2.97)$$

The definition of discrete-time white noise shows that it does not have any correlation with itself except at the present time. If $X(k)$ is a discrete-time white noise process, then the RV $X(n)$ is uncorrelated with $X(m)$ unless $n = m$. This shows that the power of a discrete-time white noise process is equal at all frequencies:

$$S_X(\omega) = R_X(0) \quad \text{for all } \omega \in [-\pi, \pi] \quad (2.98)$$

For a continuous-time random process, white noise is defined similarly. White noise has equal power at all frequencies (like white light):

$$S_X(\omega) = R_X(0) \quad \text{for all } \omega \quad (2.99)$$

Substituting this expression for $S_X(\omega)$ into Equation (2.92), we see that for continuous-time white noise

$$R_X(\tau) = R_X(0) \delta(\tau) \quad (2.100)$$

where $\delta(\tau)$ is the continuous-time impulse function. That is, $\delta(\tau)$ is a function that is zero everywhere except at $\tau = 0$; it has a width of 0, a height of ∞ , and an area of 1. Continuous-time white noise is not something that occurs in the real world because it has infinite power, as seen by comparing Equations (2.93) and (2.99). Nevertheless, many continuous-time processes approximate white noise and are useful in mathematical analyses of signals and systems.

■ EXAMPLE 2.14

Suppose that a zero-mean stationary stochastic process has the autocorrelation function

$$R_X(\tau) = \sigma^2 e^{-\beta|\tau|} \quad (2.101)$$

where β is a positive real number. The power spectrum is computed from Equation (2.92) as

$$\begin{aligned} S_X(\omega) &= \int_{-\infty}^{\infty} \sigma^2 e^{-\beta|\tau|} e^{-j\omega\tau} d\tau \\ &= \int_{-\infty}^0 \sigma^2 e^{(\beta-j\omega)\tau} d\tau + \int_0^{\infty} \sigma^2 e^{-(\beta+j\omega)\tau} d\tau \\ &= \frac{\sigma^2}{\beta-j\omega} + \frac{\sigma^2}{\beta+j\omega} \\ &= \frac{2\sigma^2\beta}{\omega^2 + \beta^2} \end{aligned} \quad (2.102)$$

The variance of the stochastic process is computed as

$$\begin{aligned} E[X^2(t)] &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2\sigma^2\beta}{\omega^2 + \beta^2} d\omega \\ &= \frac{\sigma^2\beta}{\pi} \left[\frac{1}{\beta} \tan^{-1} \frac{\omega}{\beta} \right]_{-\infty}^{\infty} \\ &= \sigma^2 \\ &= R_X(0) \end{aligned} \quad (2.103)$$

▽▽▽

2.7 SIMULATING CORRELATED NOISE

In optimal filtering research and experiments, we often have to simulate correlated white noise. That is, we need to create random vectors whose elements are correlated with each other according to some predefined covariance matrix. In this section, we will present one way of accomplishing this.

Suppose we want to generate an n -element random vector w that has zero mean and covariance Q :

$$Q = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{1n} & \cdots & \sigma_n^2 \end{bmatrix} \quad (2.104)$$

Since Q is a covariance matrix, we know that all of its eigenvalues are real and nonnegative. We can therefore denote its eigenvalues as μ_k^2 :

$$\lambda(Q) = \mu_k^2 \quad (k = 1, \dots, n) \quad (2.105)$$

Suppose the eigenvectors of Q are found to be d_1, \dots, d_n . Augment the d_i vectors together to obtain an $n \times n$ matrix D . Since Q is symmetric, we can always choose

the eigenvectors such that D is orthogonal, that is, $D^{-1} = D^T$. We therefore obtain the Jordan form decomposition of Q as

$$Q = D\hat{Q}D^T \quad (2.106)$$

where \hat{Q} is the diagonal matrix of the eigenvalues of Q . That is,

$$\hat{Q} = \text{diag}(\mu_1^2, \dots, \mu_n^2) \quad (2.107)$$

Now we define the random vector v as $v = D^{-1}w$, so that $w = Dv$. Therefore,

$$\begin{aligned} E(vv^T) &= E(D^T w w^T D) \\ &= D^T Q D \\ &= \hat{Q} \\ &= \text{diag}(\mu_1^2, \dots, \mu_n^2) \end{aligned} \quad (2.108)$$

This shows how we can generate an n -element random vector w with a covariance matrix of Q . The algorithm is given as follows.

Correlated noise simulation

1. Find the eigenvalues of Q , and denote them as μ_1^2, \dots, μ_n^2
2. Find the eigenvectors of Q , and denote them as d_1, \dots, d_n , such that

$$\begin{aligned} D &= [d_1 \ \dots \ d_n] \\ D^{-1} &= D^T \end{aligned} \quad (2.109)$$

3. For $i = 1, \dots, n$ compute the random variable $v_i = \mu_i r_i$, where each r_i is an independent random number with a variance of 1 (unity variance).
4. Set $w = Dv$.

2.8 SUMMARY

In this chapter, we have reviewed the basic concepts of probability, random variables, and stochastic processes. The probability of some event occurring is simply and intuitively defined as the number of times the event occurs divided by the number of chances the event has to occur. A random variable (RV) is a variable whose value is not certain, but is governed by the laws of probability. For example, your score on the test for this chapter is not deterministic, but is a random variable. Your *actual* score, after you take the test, will be a specific, deterministic number. But *before* you take the test, you do not know what you will get on the test. You may suppose that you will probably get between 80% and 90% if you have a decent understanding of the material, but your actual score will be determined by random events such as your health, how well you sleep the night before, what topics the instructor decides to cover on the test versus what topics you study, what the traffic was like on the way to school, the mood of the instructor when she grades the test, and so on. A stochastic process is a random variable that changes with time,

such as your performance on all of the quizzes and homework assignments for this course. The expected value of your test grades may be constant throughout the duration of the course if you are a consistent person, or it may increase if you tend to study harder as the course progresses, or it may decrease if you tend to study less as the course progresses. Probability, random variables, stochastic processes, and related topics form a huge area of study that we have only touched on in this chapter. Additional information on these topics can be found in many textbooks, including [Pap02, Pee01]. A study of these topics will allow a student to delve into many practical engineering subjects, including control and estimation theory, signal processing, and communications theory.

PROBLEMS

Written exercises

2.1 What is the 0th moment of an RV? What is the 0th central moment of an RV?

2.2 Suppose a deck of 52 cards is randomly divided into four piles of 13 cards each. Find the probability that each pile contains exactly one ace [Gre01].

2.3 Determine the value of a in the function

$$f_X(x) = \begin{cases} ax(1-x) & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

so that $f_X(x)$ is a valid probability density function [Lie67].

2.4 Determine the value of a in the function

$$f_X(x) = \frac{a}{e^x + e^{-x}}$$

so that $f_X(x)$ is a valid probability density function. What is the probability that $|X| \leq 1$?

2.5 The probability density function of an exponentially distributed random variable is defined as follows.

$$f_X(x) = \begin{cases} ae^{-ax} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

where $a \geq 0$.

- a) Find the probability distribution function of an exponentially distributed random variable.
- b) Find the mean of an exponentially distributed random variable.
- c) Find the second moment of an exponentially distributed random variable.
- d) Find the variance of an exponentially distributed random variable.
- e) What is the probability that an exponentially distributed random variable takes on a value within one standard deviation of its mean?

2.6 Derive an expression for the skew of a random variable as a function of its first, second, and third moments.

2.7 Consider the following probability density function:

$$f_X(x) = \frac{ab}{b^2 + x^2}, \quad b > 0$$

- a) Determine the value of a in the so that $f_X(x)$ is a valid probability density function. (The correct value of a makes $f_X(x)$ a Cauchy pdf.)
- b) Find the mean of a Cauchy random variable.

2.8 Consider two zero-mean uncorrelated random variables W and V with standard deviations σ_w and σ_v , respectively. What is the standard deviation of the random variable $X = W + V$?

2.9 Consider two scalar RVs X and Y .

- a) Prove that if X and Y are independent, then their correlation coefficient $\rho = 0$.
- b) Find an example of two RVs that are not independent but that have a correlation coefficient of zero.
- c) Prove that if Y is a linear function of X then $\rho = \pm 1$.

2.10 Consider the following function [Lie67].

$$f_{XY}(x, y) = \begin{cases} ae^{-2x}e^{-3y} & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

- a) Find the value of a so that $f_{XY}(x, y)$ is a valid joint probability density function.
- b) Calculate \bar{x} and \bar{y} .
- c) Calculate $E(X^2)$, $E(Y^2)$, and $E(XY)$.
- d) Calculate the autocorrelation matrix of the random vector $[X \ Y]^T$.
- e) Calculate the variance σ_x^2 , the variance σ_y^2 , and the covariance C_{XY} .
- f) Calculate the autocovariance matrix of the random vector $[X \ Y]^T$.
- g) Calculate the correlation coefficient between X and Y .

2.11 A stochastic process has the autocorrelation $R_X(\tau) = Ae^{-k|\tau|}$, where A and k are positive constants.

- a) What is the power spectrum of the stochastic process?
- b) What is the total power of the stochastic process?
- c) What value of k results in half of the total power residing in frequencies less than 1 Hz?

2.12 Suppose X is a random variable, and $Y(t) = X \cos t$ is a stochastic process.

- a) Find the expected value of $Y(t)$.
- b) Find $A[Y(t)]$, the time average of $Y(t)$.
- c) Under what condition is $\bar{y}(t) = A[Y(t)]$?

2.13 Consider the equation $Z = X + V$. The pdf's of X and B are given in Figure 2.5.

- a) Plot the pdf of $(Z|X)$ as a function of X for $Z = 0.5$.

- b) Given $Z = 0.5$, what is conditional expectation of X ? What is the most probable value of X ? What is the median value of X ?

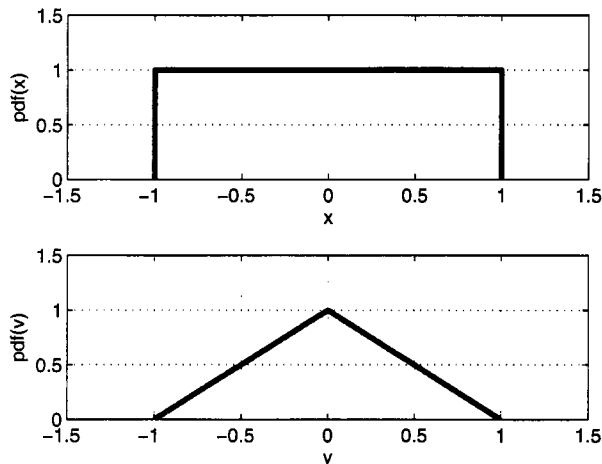


Figure 2.5 pdf's for Problem 2.13 [Sch73].

- 2.14** The temperature at noon in London is a stochastic process. Is it ergodic?

Computer exercises

2.15 Generate $N = 50$ independent random numbers, each uniformly distributed between 0 and 1. Plot a histogram of the random numbers using 10 bins. What is the sample mean and standard deviation of the numbers that you generated? What would you expect to see for the mean and standard deviation (i.e., what are the theoretical mean and standard deviation)? Repeat for $N = 500$ and $N = 5,000$ random numbers. What changes in the histogram do you see as N increases?

2.16 Generate 10,000 samples of $(x_1 + x_2)/2$, where each x_i is a random number uniformly distributed on $[-1/2, +1/2]$. Plot the 50-bin histogram. Repeat for $(x_1 + x_2 + x_3 + x_4)/4$. Describe the difference between the two histograms.