Chapter 1

Discussion of Probability and Stochastic Processes

Introduction

The purpose of this book is to present some particular topics from the theory of stochastic processes which have found applications in control and communications engineering. The book has been written on the assumption that the reader has already had an introductory course in probability theory. Nevertheless, for a variety of reasons it seems appropriate and useful to begin with a review of that subject.

In this chapter we provide a review of the main ideas from probability theory that will be needed in understanding the material in this book. Beyond that, we will introduce one or two ideas which will probably be new to the reader, such as the Hilbert space of second-order random variables, that also will be handy to have available. Finally, after we define some terms and develop some concepts, we will explain what we hope the student will acquire from studying the material in this book, and provide a brief survey of the task to be undertaken.

In order to do that, we will provide a tentative definition of the term "stochastic process," as well as a brief discussion of certain kinds of stochastic processes which will be encountered again subsequently. The end of the chapter also contains a short statement explaining why the book has been written the way it has.

Probability

It is widely agreed that a good way to study probability theory is to base it on set theory. We will approach the subject from that standpoint. The term

"set" is, in very rigorous treatments, considered to be an undefined concept which includes certain properties that are assumed in the initial axioms upon which the whole subject is based. Intuitively, a set is a collection of objects. In probability theory, these "objects" are *elementary events*. In set theory, the set of all the objects with which one intends to deal is taken as the *universal set*. In probability theory, the universal set is called the *sample space*.

Suppose one does an experiment in which the element of randomness is known to play a role. For example, conduct a survey by selecting some category of people and asking them questions, or make repeated measurements of some physical variable under circumstances where experimental error is known not to be negligible. Such an experiment is sometimes called a *random experiment*. It is not the structure of the experiment that is random; instead, randomness refers to the fact that the outcome cannot be predicted precisely in advance.

The *statistics* of the experiment refers, at the most primitive level, simply to the data itself. On a more refined level, "statistics" also refers to certain properties the data is found to have after subjecting it to some numerical processing. Probability theory is used to analyze such a random experiment. It is used to decide what kind of numerical processing is appropriate for the data and what kind of statements one can make with confidence concerning the statistics. Even more basically, probability theory is used to determine how the experiment should be structured so that one *can* make meaningful statements with confidence.

In performing such an analysis using probability theory, it turns out to be a disadvantage to have a sample space that is too large or too small. Therefore, the choice of sample space is usually tailored to the experiment in question. For example, suppose the experiment is to flip a coin 10 times, and record the outcome of each flip, that is, whether it is heads or tails. A sample space with only two points in it, heads and tails, is too small and is actually not useful. A sample space with infinitely many points in it is certainly large enough. The problem is, it is so large as to be unwieldy, and it may lead one into mathematical distress of a kind that one prefers to avoid if there be a way of avoiding it.

The sample space for the above experiment which turns out to be "just right" is the set of all *binary sequences of length 10*. There are $2^{10} = 1024$ of these, so this sample space contains 1024 points. Each point is an "elementary event," that is, a complete sequence of 10 flips. A single flip is *not* an elementary event.

In doing mathematical probability theory this way, a numerical probability would *first* be assigned to *each elementary event* (each sequence of length 10). The value of the probability assigned to each event must be a

real number between 0 and 1, and the sum of the values over all 1024 points of the sample space must be exactly 1.

At this juncture, we can look at various subsets of the sample space, for example, the subset consisting of all sequences having heads occur on the first flip. The sum of the values of probability over all of the points in this subset is, *by definition*, the probability of getting a head on the first flip. If that number agrees with what you intuitively feel ought to be the case, then you may say that your coin-flipping model is *realistic*. On the other hand, if that is not the value that you think the event of getting a head on the first flip should have, then you must change the probabilities assigned to the elementary events until things come out the way you want them to.

Probability theory will show you how to make calculations from your mathematical model concerning the probabilities of various events. It is up to *you* to take the responsibility for deciding whether or not the model is realistic. If you test it in situations where the correct answer is already known, and the model gives you the correct answer there, then you may feel confident in trusting it in situations where the answer is unknown.

Let us now give some precise mathematical definitions. The fundamental entity that we require in order to use probability theory is a *probability trio* $(\Omega, \mathscr{A}, P)$. The first member of the trio, Ω, is the *sample space*, which may be either finite, countably infinite, or uncountably infinite. The second member of the trio, $\mathscr{A}$, is the *algebra of admissible subsets of* Ω, also called the *algebra of events*. The third member of the trio, P, is the *probability measure* defined on $\mathscr{A}$. That is, P is a set function. Its argument is one of the sets that belongs to $\mathscr{A}$, and its value is a real number between 0 and 1.

If Ω is a finite set, then $\mathscr{A}$ is simply the collection of *all* subsets of Ω, the so-called *power set* 2^{Ω}. If Ω is an infinite set, it is not possible in general to assign a probability to every one of its subsets in a consistent way without encountering mathematical difficulties. Therefore, the family of subsets of Ω to which probabilities are assigned has to be specified. That is what $\mathscr{A}$ is. Its members obey the rules of Boolean algebra with respect to the operations of union, intersection, and complement.

With these agreements in force, the only conditions that the set function P must satisfy in order to be a probability measure are the following:

1. $P(\varnothing) = 0$ where $\varnothing$ = empty set
2. $P(\Omega) = 1$
3. $P(A) \geq 0$ for every A in $\mathscr{A}$
4. If $A_1, A_2, \ldots$ are disjoint members of $\mathscr{A}$, then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$$

Random Variables

In addition to the function P defined on $\mathscr{A}$, we also consider functions defined on Ω itself. Any such function is called a *random variable*. If the value of the function is a real number, it is called a real random variable; if the value is a complex number it is called a complex-valued random variable; if the value is a vector in R^n, it is called a vector-valued random variable; and so on. It is customary to abbreviate "random variable" by r.v.

If the set Ω is infinite, then in order to avoid mathematical distress we have to ban certain pathological functions. It is very unlikely such a function would arise in most applications, but we will include this restriction for the sake of precision. Let us explain it further.

The class of admissible random variables must agree with our algebra of admissible sets. We will explain what "agree" means for real r.v.'s; the extension to more general r.v.'s is a technicality. If $X(\omega)$ is a real r.v., then we want to discuss the probability that the value of X falls in some interval I of the real line. In order to do that, we have to be dealing with an event. Therefore, define

$$X^{-1}(I) = \{\omega \in \Omega\colon X(\omega) \in I\} \tag{1}$$

The symbol $\in$ means "belongs to." It suffices for this condition to consider only the class of semi-infinite intervals of the form $I = (-\infty, a]$, for every real number a. If for each a, the set $X^{-1}(I)$ is a member of $\mathscr{A}$, then X is an admissible r.v.

Under those circumstances, we are assured that the probability $P\{-\infty < X \le a\}$ of the event that X is less than or equal to a is well defined. We give this probability a special name. Since it is a function of the parameter a, we call it the *distribution function for the r.v. X*. It is denoted by $F_X(a)$. In symbols:

$$F_X(a) = P\{-\infty < X \le a\} \tag{2}$$

Under appropriate circumstances, the distribution function $F_X(a)$ turns out to be differentiable with respect to the parameter a. This will happen only when the sample space Ω is uncountably infinite. In those cases it is convenient to work with the *probability density function*, defined as the derivative of F_X. It has become a common practice to use the same letter for the argument of this density function as is used to designate the random variable itself. Although this system may be used without confusion by those proficient in the subject, for students trying to master the fundamentals it is misleading and confusing. In this book we will always use a capital letter for random variables. The parameter in the density function will then be the corresponding lowercase letter.

DEFINITION. Let X be a real random variable having a probability distribution F_X which is differentiable. Denote the derivative by f_X. Then we call f_X the *probability density function for the r.v. X*. In symbols:

$$f_X(x) = \frac{d}{dx}F_X(x) \tag{3}$$

The values of F_X are probabilities, but the values of f_X are not. Probabilities are found by integrating f_X, for example:

$$P\{a \le X \le b\} = \int_a^b f_X(x)\,dx \tag{4}$$

It follows directly from (2) that the distribution function F_X for any r.v. X possesses the following four properties:

1. F_X is nondecreasing: $a < b$ implies $F_X(a) \le F_X(b)$
2. $\lim_{x \to +\infty} F_X(x) = 1$
3. $\lim_{x \to -\infty} F_X(x) = 0$
4. F_X is continuous from the right, that is, at any discontinuity F_X assumes the upper value.

If F_X is piecewise constant, that is, a staircase function consisting of only finite jumps and constant segments, then X is called a discrete r.v. If F_X has no discontinuities whatsoever, then X is called a *continuous* r.v. A general r.v. is sometimes called *mixed*.

Strictly speaking, only continuous r.v.'s with F_X differentiable can possess density functions, although by resorting to the use of δ functions, which is common in engineering practice, even a discrete r.v. can be assigned a density.

Suppose X is a discrete r.v. which assumes only a finite set of possible values $a_1, a_2, \ldots, a_n$, with respective probabilities $p_1, p_2, \ldots, p_n$. Intuitively, we may say that X can be expected to have value a_k a fraction p_k of the time. If we make many different observations of X and average the results, then as the number of observations becomes infinite the sample average will approach the number

$$\mu = \sum_{k=1}^{n} a_k p_k \tag{5}$$

In (5) we have written μ as a sum over the *range of* X, that is, the set of values assumed by X. Conceptually, it is valuable to realize that this same quantity could also be computed by a sum over the sample space Ω, specifically

$$\mu = \sum_{\omega \in \Omega} X(\omega) P\{\omega \in \Omega\colon X(\omega) = a_k\} \tag{6}$$

The summation in (6) is accomplished by partitioning Ω into disjoint subsets $A_1, A_2, \ldots, A_n$, such that for each k, A_k is the set of ω points for which $X(\omega)$ assumes the same value a_k.

When X is a continuous r.v., the definition (5) generalizes to

$$\mu = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{7}$$

The expression (6) generalizes into the Lebesgue integral, as defined in measure theory. A discussion of that is beyond the scope of this book.

The quantity given by (5), (6), or (7) is called the *mean* or *expected value* of X. In rigorous treatments, the most satisfactory way of introducing the expected value operator is to base it on a precise version of (6), which we have here written in a symbolic form to try to suggest the underlying concept.

Since we will mainly be concerned with r.v.'s possessing density functions, we will henceforth take (7) as the definition of the mean, without further comment.

Higher moments are defined analogously, whenever the integrals exist:

$$\mu_n = \int_{-\infty}^{\infty} x^n f_X(x)\, dx \tag{8}$$

When considerable work has to be done involving moments, it is useful to make use of the properties of the *characteristic function* $M(u)$, which is just the Fourier transform of the density:

$$M(u) = \int_{-\infty}^{\infty} e^{iux} f(x)\, dx \tag{9}$$

When the moment μ_n exits, it may be found by the formula

$$\mu_n = (-i)^n \frac{d^n}{du^n} M(u)\bigg|_{u=0} \tag{10}$$

If the characteristic function is known, then the density may be recovered

by taking the inverse Fourier transform:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iux} M(u)\, du \tag{11}$$

Be careful to note that in the above definitions, the word "inverse," the factor $1/2\pi$, and the minus sign in the exponent have been permuted with respect to the way they are commonly arranged in defining Fourier transforms of functions of time. In dealing with such permutations, the thing that always remains unchanged is the fact that

$$f(x) = \int_{-\infty}^{\infty} \delta(x - x') f(x')\, dx' \tag{12}$$

Now, the δ function can always be represented by either

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{iu(x-x')}\, du \tag{13}$$

or by

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iu(x-x')}\, du \tag{14}$$

Depending upon what is called the forward transform and what is called the inverse, it must always be true that either

$$f = \mathscr{F}^{-1}[\mathscr{F}[f]] \tag{15}$$

or

$$f = \mathscr{F}[\mathscr{F}^{-1}[f]] \tag{16}$$

Whichever applies, (15) or (16) must reduce to (12) when the appropriate representation of the δ function is used from (13) or (14).

Fourier transforms occur in this text not only in connection with probability density and characteristic functions, but also in connection with autocovariance and, to be defined later on, power spectral density functions. Since the definitions of these objects do vary from one textbook to another, it is hoped that the above discussion will help dispel some of the resultant confusion. Any variation in the definition is permissible as long as one remains consistent with their own definition and with the above principles.

A new r.v. Y can be generated from an existing r.v. X by making Y a function of X:

$$Y = g(X) \tag{17}$$

Applying the definition of the mean of Y leads us to define, in general, the *expected value of the function* $g(X)$ by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \tag{18}$$

whenever the integral exists.

It is also possible to have several r.v.'s $X_1, X_2, \ldots, X_n$ defined on the same underlying probability trio $(\Omega, \mathscr{A}, P)$, where there is no functional relationship like (17) connecting one r.v. to another. To handle this situation, one uses the *joint cumulative distribution function* $F_{X_1, X_2, \ldots, X_n}(a_1, a_2, \ldots, a_n)$, defined by

$$F_{X_1, X_2, \ldots, X_n}(a_1, a_2, \ldots, a_n)$$

$$= P\{-\infty < X_1 \leq a_1, -\infty < X_2 \leq a_2, \ldots, -\infty < X_n \leq a_n\} \tag{19}$$

If this function is jointly differentiable with respect to all of its arguments, then the *joint density function* $f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)$ is defined by

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)$$

$$= \frac{\partial^n}{\partial x_1, \partial x_2, \ldots, \partial x_n} F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \tag{20}$$

In this case it is usually expedient to introduce the vector-valued random variable

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \tag{21}$$

When this is done, the joint density defined in (20) might be denoted simply $f_{\mathbf{X}}(\mathbf{x})$.

Most of the features of one-dimensional densities can be extended in an obvious way to joint densities. For example, by use of the n-dimensional Fourier transform, a joint characteristic function is defined as a generalization of (9).

Independence and Conditional Probability

Given a probability trio $(\Omega, \mathscr{A}, P)$, let A and B be two members of $\mathscr{A}$. If the measure P assigns probabilities in such a way that

$$P(A \cap B) = P(A)P(B) \tag{22}$$

then we say that the events A and B are *independent*.

Whether A and B are independent or not, if $P(B) > 0$ it is customary to define the ratio

$$\frac{P(A \cap B)}{P(B)} = P(A|B) \tag{23}$$

and $P(A|B)$ is called the *conditional probability of A given B*. In terms of it, the condition (22) for independence may be written

$$P(A|B) = P(A) \tag{24}$$

which says, knowledge of whether or not the event B has occurred has no influence upon the probability that event A occurs.

Suppose X and Y are r.v.'s defined on the same trio $(\Omega, \mathscr{A}, P)$. Assume they are continuous r.v.'s, and let their joint density be $f_{XY}(x, y)$. The two one-dimensional densities for each r.v. considered by itself, denoted respectively $f_X(x)$ and $f_Y(y)$, are called *marginal* densities. They can each be found by *marginal integration*:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \tag{25}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \tag{26}$$

Two r.v.'s that possess a joint density function are *independent if and only if*

$$f_{XY}(x, y) = f_X(x) f_Y(y) \tag{27}$$

The ratio

$$\frac{f_{XY}(x, y)}{f_Y(y)} = f_{X|Y}(x|y) \tag{28}$$

is called the *conditional density for X, given that* $Y = y$.

For example if a and b are two real numbers, then

$$\int_a^b f_{X|Y}(x|y)\, dx = P(\{a \leq X \leq b\}|\{Y = y\}) \tag{29}$$

If Y is a continuous r.v., then the event $\{Y = y\}$ has probability zero. The conditional probability on the right-hand side of (29) therefore cannot be defined by a simple straightforward application of the definition (23). The conditional probability in (29) can, nevertheless, be defined in a way that is totally satisfactory from a rigorous standpoint, but it requires use of a technical device from measure theory, called a sigma-field, which is beyond the scope of this text.

Let $X_1, X_2, \ldots, X_n$ all be continuous r.v.'s defined on the same trio $(\Omega, \mathscr{A}, P)$, and suppose they have a joint density $f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)$. Let m be an integer such that $1 < m < n$. The conditional density for $X_{m+1}, X_{m+2}, \ldots, X_n$, given $X_1, X_2, \ldots, X_m$, denoted $f_{X_{m+1}, \ldots, X_n | X_1, \ldots, X_m}(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m)$ is defined as the ratio

$$f_{X_{m+1}, \ldots, X_n | X_1, \ldots, X_m}(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = \frac{f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)}{f_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m)} \tag{30}$$

The denominator of (30) is found by marginal integration:

$$f_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-m} f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)\, dx_{m+1}, \ldots, dx_n \tag{31}$$

The Hilbert Space of Second-Order Random Variables

A *Hilbert Space* is a vector space equipped with an inner product and a norm that is derived from the inner product. Finite-dimensional Hilbert space is just an abstraction and generalization of finite-dimensional Euclidean space. Infinite-dimensional Hilbert space is the extension of this concept to an infinite number of dimensions. In that case the definition must also be expanded to include the attribute "complete." "Complete" means that every infinite sequence of vectors drawn from the space, which is Cauchy in the sense of the norm, converges to a limit vector that also belongs to the space.

We will discuss the finite-dimensional case first. Suppose we have a set of n random variables $X_1, X_2, \ldots, X_n$, each of which has finite second moment:

$$EX_k^2 < \infty, \qquad k = 1, 2, \ldots, n \tag{32}$$

An r.v. X_k that obeys (32) is called a *second-order* r.v.

Our first task is to introduce the concepts of linear independence and statistical independence.

DEFINITION. The set of second-order r.v.'s $X_1, X_2, \ldots, X_n$ is called *linearly independent* if and only if the equation

$$c_1 X_1 + c_2 X_2 + \cdots + c_n X_n = 0 \quad \text{implies} \quad c_1 = c_2 = \cdots = c_n = 0 \tag{33}$$

Let $F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)$ be the joint cumulative distribution of $X_1, X_2, \ldots, X_n$, that is,

$$F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = P\{-\infty < X_k \le x_k,\ k = 1, 2, \ldots, n\}$$

Let $F_{X_k}(x_k)$, $k = 1, 2, \ldots, n$ be the marginal cumulative distribution function for each r.v. X_k, $k = 1, 2, \ldots, n$.

DEFINITION. The set of r.v.'s $X_1, X_2, \ldots, X_n$ is called *statistically independent* if and only if

$$F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} F_{X_k}(x_k) \tag{34}$$

This is the usual definition of mutual independence, generalizing (27). The r.v.'s $X_1, \ldots, X_n$ need not have finite second moments in order for this definition to be usable. Also, there is no assumption that the means $E[X_n]$ are zero. However, if they all have both zero mean and finite second moment, then statistical independence implies linear independence, but not conversely.

DEFINITION. Let X_j and X_k be second-order r.v.'s. Their *inner product*, denoted $\langle X_j, X_k \rangle$, is defined as

$$\langle X_j, X_k \rangle = E[X_j X_k] \tag{35}$$

DEFINITION. Let X be a second-order r.v. Its *norm*, denoted $\|X\|$, is defined as

$$\|X\| = \sqrt{E[X^2]} \tag{36}$$

Now let $X_1, X_2, \ldots, X_n$ be any set of n linearly independent second-order r.v.'s Consider the set of all possible liner combinations of these r.v.'s, that

is, all other r.v.'s of the form

$$Y = \sum_{k=1}^{n} c_k X_k \tag{37}$$

We then have

$$Y^2 = \left(\sum_{k=1}^{n} c_k X_k \right)^2$$

$$= \sum_{k=1}^{n} \sum_{j=1}^{n} c_k c_j X_k X_j \tag{38}$$

where we have written the product of two single sums as a double summation by changing the dummy index.

One of the problems at the end of the chapter is to prove the Schwarz inequality

$$|\langle X, Y \rangle| \leq \|X\| \, \|Y\| \tag{39}$$

Taking the expected value of both sides of (38) yields

$$\|Y\|^2 = E[Y^2] = \sum_{h=1}^{n} \sum_{j=1}^{n} c_k c_j E[X_k X_j]$$

$$= \sum_{k-1}^{n} \sum_{j=1}^{n} c_k c_j \langle X_k, X_j \rangle$$

$$\leq \sum_{k=1}^{n} \sum_{j=1}^{n} |c_k| \, |c_j| \, |\langle X_k, X_j \rangle|$$

$$\leq \sum_{k=1}^{n} \sum_{j=1}^{n} |c_k| \, |c_j| \, \|X_k\| \, \|X_j\|$$

$$= \left(\sum_{k=1}^{n} |c_k| \, \|X_k\| \right)^2 \tag{40}$$

Thus $\|Y\|^2 \leq (\Sigma_{k=1}^{n} |c_k| \, \|X_k\|)^2 < \infty$ provided $|c_k| < \infty$, $k = 1, 2, \ldots, n$, so every such r.v. of the form (37) is a second-order r.v. The set of all such r.v.'s Y is our Hilbert space.

We may construct an orthogonal basis for the space $V_1, V_2, \ldots, V_n$ by applying the Gram–Schmidt procedure to $X_1, X_2, \ldots, X_n$:

$$V_1 = X_1$$

$$V_2 = X_2 - \frac{\langle X_2, V_1 \rangle}{\langle V_1, V_1 \rangle} V_1 \tag{41a}$$

and for arbitrary k,

$$V_k = X_k - \sum_{j=1}^{k-1} \frac{\langle X_k, V_j \rangle}{\langle V_j, V_j \rangle} V_j \tag{41b}$$

If the original set $\{X_1, X_2, \ldots, X_n\}$ were not in fact all linearly independent, then the above procedure will simply return zero for the corresponding V_k, whenever X_k is not linearly independent of $\{X_1, \ldots, X_{k-1}\}$. In that case, merely continue with the procedure, deleting V_k. When finished, the resulting set $\{V_1, V_2, \ldots, V_m\}$ for some $m < n$ will be an orthogonal basis, and m will be the dimension of the space spanned by $\{X_1, X_2, \ldots, X_n\}$.

The same procedure can be used in the case of infinitely many elements $X_1, X_2, \ldots,$ and in principle could be used to determine whether the HIlbert space is finite-dimensional or infinite-dimensional. The possibility that we may be dealing with an *infinite* number of r.v.'s all defined on the same probability trio $(\Omega, \mathscr{A}, P)$, which are mutually linearly independent, will be very important to us for the remainder of this book, because it is a fundamental concept in the theory of *random processes*, to which we now turn.

Random Processes

Having raised the possibility of an infinite family of r.v.'s all defined on the same sample space, we now formalize the concept.

DEFINITION. A *random process* (equivalently: *stochastic process*) is a family $\{X_t: t \in T\}$ of random variables, all defined on the same probability trio $(\Omega, \mathscr{A}, P)$. The set T is the *parameter set* of the random process.

In this book, all of our random processes will either be real valued, complex valued, or vector valued (vector in R^n). If the parameter set T is the set of integers or a subset thereof, the process is called a *discrete parameter* process. If T is a subset of the real line, then the process is called a *continuous parameter* process.

In general, other possibilities exist for the space in which the r.v.'s take values (sometimes called the *state space* of the process) and for the parameter set. The ones mentioned above are the only ones used in this book.

In Chapters 7 and 10 we will discuss a particular category of random processes known as Markov processes. It is appropriate to provide the relevant definition here.

DEFINITION. The random process $\{X_t: t \in T\}$ is called a *Markov process* provided the following circumstances hold:

Let S be any subset of the state space. Let t_f (the future time) and t_p (the present time) be any two elements of T with $t_f > t_p$. Let Q (the set of past times) be any subset of T containing t_p such that, for every $t \in Q$, $t_p \geq t$. Then

$$P\{X_{t_f} \in S | X_t, t \in Q\} = P\{X_{t_f} \in S | X_{t_p}\} \tag{42}$$

In words, this definition says that a Markov process is one having the property that, given the present state of the process, the future becomes conditionally independent of the past. To illustrate this further, suppose $\{X_t: t \in T\}$ is a Markov process such that all of the r.v.'s X_t are continuous r.v.'s. Let $t_1 < t_2 < \cdots < t_{n-1} < t_n$ be a set of points in T. Then we may consider the joint density function $f_{X_{t_1}, X_{t_2}, \ldots, X_{t_n}}(x_1, x_2, \ldots, x_n)$ and the associated conditional density function (with $1 < m < n$)

$$f_{X_{t_{m+1}}, X_{t_{m+2}}, \ldots, X_{t_n} | X_{t_1}, X_{t_2}, \ldots, X_{t_m}}(x_{m+1}, x_{m+2}, \ldots, x_n | x_1, x_2, \ldots, x_m)$$

as defined in (30). Let t_m play the role of the present time t_p in the preceding definition. Consider $t_{m+1}, \ldots, t_n$ as future times, and $t_1, \ldots, t_{m-1}$ as past times. Then, for *any* set $\{t_1, t_2, \ldots, t_n\}$ chosen from T with the preceding properties, the Markov nature of $\{X_t: t \in T\}$ means that the following equation is *identically true*:

$$f_{X_{t_{m+1}}, \ldots, X_{t_n} | X_{t_1}, \ldots, X_{t_m}}(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m)$$

$$= f_{X_{t_{m+1}}, \ldots, X_{t_n} | X_{t_m}}(x_{m+1}, \ldots, x_n | x_m) \tag{43}$$

That completes our consideration of Markov processes in this chapter. We return to the discussion of general random processes.

Since the qualification "$t \in T$" is always understood, henceforth in this book we shall simply write a random process as $\{X_t\}$. Again let

$t_1 < t_2 < \cdots < t_n$ be an ordered set of parameter points (usually the parameter will be interpreted as time). Consider the random vector $\mathbf{X}$ of the values $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$, as in (21), and denote the joint density function by $f_{\mathbf{X}}(\mathbf{x})$.

Let $\mathbf{C}$ be an $n \times n$ matrix of real numbers, which is symmetric and positive definite. Let $\boldsymbol{\mu}$ be an n-vector of real numbers. We will provide a fuller discussion of Gaussian distributions in Chapters 2 and 3, but it is appropriate to introduce the following definition now:

DEFINITION. The random process $\{X_t\}$ is called a *Gaussian* process provided that for any selection of the ordered set $\{t_1, t_2, \ldots, t_n\}$ from T, for any integer n, there exists a positive definite $n \times n$ matrix $\mathbf{C}$ and an n-vector $\boldsymbol{\mu}$ such that the joint density function for the random vector $\mathbf{X}$ pertinent to these time points is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{44}$$

Much of the remainder of this book is devoted to the consideration of Gaussian processes. In (44), the notation $|\mathbf{C}|$ means the determinant of the matrix $\mathbf{C}$.

Two very important entities associated with any Gaussian process $\{X_t\}$ are its *mean* μ_t,

$$\mu_t = E\{X_t\} \tag{45}$$

and its *autocovariance* c_{ts},

$$c_{ts} = E\{[X_t - \mu_t][X_s - \mu_s]\} \tag{46}$$

both defined for all t, s in T. Conversely, we show later that knowledge of these two entities is sufficient to characterize completely a Gaussian process.

Be careful not to jump to the conclusion that, just because the mean μ_t and covariance c_{ts} are given for some process $\{X_t\}$, it necessarily follows that $\{X_t\}$ is Gaussian. This conclusion is false. There are many possible distribution laws having the same first two moments as a particular Gaussian distribution, but for which the higher moments are entirely different. As a specific example, the distribution could be bimodal. There is no way to tell whether a distribution is unimodal, bimodal, or multimodal by looking only at the first two moments.

In the applications for the material presented in this book, one is usually dealing with a physical situation in which there is an element of randomness and where important features of the situation are investigated by examining a collection of measurements that emerge sequentially in time. That is the sort of situation one would attempt to model by means of a random process. After constructing such a model one would hope that mathematical analysis of the model will lead one to insights that are valid and pertinent to the actual physical situation.

In the selection and construction of such a model, there is always the question of how detailed and explicit to make it. In the present context, this question would be relevant to the selection of an appropriate model when all one knows about some process $\{X_t\}$ are data that are equivalent to the first two moments, or equivalently, to the mean μ_t and the autocovariance c_{ts}. The decision that has to be made in that case is whether to commit oneself to a specific probability distribution law or whether to leave that feature open.

The disadvantage of choosing a model based on an explicitly specified distribution (e.g., Gaussian) is that one has made the model more detailed and explicit than the situation, or the data available, actually warrants. In turn, this leads one into a false sense of confidence to make inferences and extrapolate results far beyond what is justified by the knowledge available. This is the peril of overspecificity.

On the other hand, if one chooses to duck the issue and assume no specific probability law, then the only calculations that can ever be made are those based specifically (in the case under consideration) on the first two moments. Means and covariances of various r.v.'s can be calculated, but it is never possible to calculate probabilities of events.

If the only operations ever performed on the processes under consideration are linear, then it turns out that only the pertinent mean and covariance functions ever need be considered. That is, the mean, autocovariance of the output, and cross-covariance between output and input of a linear system can be calculated knowing only the mean and autocovariance of the input and the transfer function of the system. A body of theory exists for carrying out precisely such calculations for second-order processes in linear systems. In this way a stochastic system can be analyzed using only deterministic quantities. For many years, this was considered to be a great advantage.

There is a recent development that tends to reverse the situation: the widespread popularity of computer simulation. Rather than just using the computer to carry out the above-mentioned calculations based on second-order theory, the incredibly high speed of modern computers makes it

feasible both technically and economically to do so-called Monte Carlo simulations of stochastic systems. In such a simulation, the computer generates an ensemble of waveforms with the appropriate statistical characteristics. These waveforms are applied, one by one, as inputs to the system, and the resulting outputs are recorded. Statistical analysis of the resulting ensemble of outputs then permits one to make whatever calculations and inferences are appropriate to the system under consideration.

Because Monte Carlo simulation can be performed just as readily for time-varying and nonlinear systems as for time-invariant linear systems, this approach is steadily gaining wider acceptance and greater favor. The reason the use of Monte Carlo simulations reverse the preference for second-order models over specifically Gaussian models is that, in setting up the simulation, one has to adopt some specific distribution for the random numbers being generated. There are more compelling reasons to choose the Gaussian distribution than any other with the same second-order statistics.

It is quite likely that any contemporary serious worker in applied stochastic processes will become involved in computer simulations. In this work, he will find it very handy to know some techniques for working with actual Gaussian distributions, beyond the methods of second-order theory. For this reason, this book includes considerable discussion of certain features of Gaussian distributions that are likely to be useful in running computer simulations. In particular, the next three chapters focus on this material.

Starting with the simple one-dimensional Gaussian distribution in Chapter 2, we cover the multidimensional distribution in Chapter 3 and move toward random process theory by discussing sequences of finite length in Chapter 4. Finally in Chapter 5 we present the classical second-order theory for discrete time sequences, and move to continuous-time processes in Chapter 6. Chapters 7, 8, and 9 cover more advanced subjects involving continuous-time stochastic processes. Chapters 10 and 11 cover some advanced subjects involving discrete-time stochastic processes. After having completed the study of Chapters 1–6, the sets $\{7, 8, 9\}$ and $\{10, 11\}$ are independent of each other. Either set may be studied on its own, at the students' (or the instructor's) convenience and discretion.

Another recurring theme throughout this book is triangular factorization of covariance matrices, which finds its parallel in continuous time in the spectral factorization technique. In order to attempt to dispel the mystery of this topic, it is introduced in the next chapter in the familiar procedure of completing the square. Appendix 1 gives the basic pertinent theorem. The idea reaches another culmination in the final chapter of the book, where it is used to derive the Kalman filter via the concept of the innovations

process. In the discussion of the innovations process in Chapter 11, we will again meet the Gram–Schmidt orthogonalization procedure in the Hilbert space of second-order random variables.

This book not only moves upward and outward to provide the reader with an ongoing confrontation with new topics, but also periodically revisits topics already discussed, showing how they reappear in a new guise. To some extent, the organization of the book therefore resembles a spiral. We hope the reader enjoys the journey.

Problems

1. The sample space S has three elements: $S = \{s_1, s_2, s_3\}$. The set function $Q(\cdot)$ assigns numbers as follows:

$$Q(\{s_1\}) = \tfrac{1}{4} \qquad Q(\{s_3\}) = \tfrac{1}{4}$$
$$Q(\{s_2\}) = \tfrac{1}{4} \qquad Q(S) = 1$$

Is $Q(\cdot)$ a probability measure? If not, what conditions must be changed to make it one?

2. For every subset A of the real line, let $N(A)$ be the set function whose value at A is equal to the number of points in A which are positive integers. For example, if $A_0 = \{x\colon 6\tfrac{1}{2} \le x \le 7\tfrac{1}{2}\}$, then $N(A_0) = 1$.

Now let

$$A_1 = \{x\colon x \text{ is a multiple of 3 and } x \le 50\}$$
$$A_2 = \{x\colon x \text{ is a multiple of 7 and } x \le 50\}$$

a. Find $N(A_1)$, $N(A_2)$, $N(A_1 \cup A_2)$, $N(A_1 \cap A_2)$.

b. Verify that $N(A_1 \cup A_2) = N(A_1) + N(A_2) - N(A_1 \cap A_2)$.

3. A bag contains seven red balls and three green balls. A box contains five red balls and six green ones. Three balls are selected at random from the bag and are transferred to the box, after which a ball is selected at random from the box.

a. What is the probability that the ball drawn from the box is red?

b. Given that the ball from the box is red, what is the conditional probability that two or more of the balls transferred were red?

4. The bivariate r.v. (X, Y) has the joint density function

$$f_{XY}(x, y) = \begin{cases} \dfrac{x^2 - y^2}{8} e^{-x}, & 0 \le x \le \infty, -x \le y \le x \\ 0, & \text{otherwise} \end{cases}$$

Find the marginal densities $f_X(x)$ and $f_Y(y)$, and the conditional density $f_{Y|X}(y|X = x)$.

5. Let X be a Gaussian random variable having the probability density function

$$f_X(u) = (2\pi)^{-1/2} \exp\left(-\frac{u^2}{2}\right)$$

The random variable Y is defined as $Y = X^3$. Determine and plot the probability density function for Y.

6*. The stochastic process $X(t)$ is defined by

$$X(t) = \sin(at + B)$$

where a is constant and B is a random variable uniformly distributed on the interval $[0, 2\pi]$. Find the cumulative distribution function

$$F_X(u) = P\{-\infty < X(t) \le u\}$$

and the probability density function

$$f_X(u) = \frac{d}{du} F_X(u)$$

7*. The stochastic process $X(t)$ is defined by

$$X(t) = e^{tA}, \qquad 0 \le t \le \infty$$

where A is a random variable with probability density function $f_A(a)$. Find the distribution function

$$F_X(u) = P\{-\infty < X(t) \le u\}$$

and the density

$$f_X(u) = \frac{d}{du} F_X(u)$$

8. Urn #1 and urn #2 each initially contain six red, four white, and eight blue balls. At each step, one ball is selected at random from each urn, and the two balls interchange urns.

**Note:* In Problems 6 and 7, the distribution and density functions will also depend on the time t.

At time n, let the random variable X_n be the number of white balls in urn #1. Is the random sequence $X_0, X_1, X_2, \ldots, X_n, \ldots$, a Markov chain?

If you say yes, then determine the transition function $P(x, y)$ defined as $P(x, y) = P\{X_k = x | X_{k-1} = y\}$.

If you say no, explain why it's not Markov.

9. Suppose the procedure in Problem 8 is modified as follows: At each step, one ball is still selected at random from each urn. However, if it is red, then it is replaced in the *same* urn from which it was drawn. If it is blue or white, then it is placed in the opposite urn.

Repeat Problem 8 for this situation.

10. Let $\mathscr{H}$ be a Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Prove the *Schwarz inequality*: for any vectors x, y in $\mathscr{H}$, it holds that

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$$

Hint: Start from the fact that for all choices of scalars α, β, it holds that

$$\langle (\alpha x + \beta y), (\alpha x + \beta y) \rangle \geq 0$$

and choose α and β suitably.

Chapter 2

The Gaussian Distribution in One and Two Dimensions

The One-Dimensional Gaussian Distribution

A real-valued random variable X is said to possess a *Gaussian distribution* if, for real numbers a, b with $a < b$ it holds that

$$P\{a \leq X < b\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b e^{-(x-\mu)^2/2\sigma^2} \, dx \tag{1}$$

In this expression, μ and σ are parameters of the distribution. By direct calculation, one finds that

$$E\{X\} = \mu$$

$$E\{(X - \mu)^2\} = \sigma^2$$

so that μ is the mean and σ^2 is the variance of the random variable X.

The integral in (1) cannot be evaluated analytically, so in order to do numerical calculations, one must resort to tables.

The function under the integral sign in (1), specifically,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{2}$$

is called the *Gaussian probability density function*. We assume that its principal features are already familiar to the student.