CHAPTER 3

Probability theory and static models

3.1 INTRODUCTION

This chapter lays the foundation for a model of stochastic events and processes, events and processes which deterministic models cannot adequately describe. We want such a model to account explicitly for the sources of uncertainty described in Chapter 1. To do so, we will first describe random events probabilistically, and then in the next chapter we will add dynamics to the model through a study of random processes.

Probability theory [3–5, 8–10] is basically addressed to assigning probabilities to events of interest associated with the outcomes of some experiment. The fundamental context of this theory is then a general space of outcomes, called a sample space. However, it is more convenient to work in a Euclidean space, so we consider a function, or mapping, from the sample space to Euclidean space, called a random variable. The basic entity for describing probabilities and random variables becomes the probability distribution function, or its derivative, the probability density function, which will exist for the problems of interest to us. These concepts are discussed in Sections 3.2 and 3.3. Since we will eventually want to generate probability information about certain variables of interest, based upon measurements of related quantities, conditional probability densities and functions of random variables will be of primary importance and are described in Sections 3.4 and 3.5.

Having a mathematical model of probability of events, one can consider the concept of expected values of a random variable, the average value of that variable one would obtain over the entire set of possible outcomes of an experiment. Again looking ahead to estimation of variables based on measurement data, the idea of conditional expectation arises naturally. Expectations of certain functions yield moments of random variables, a set of parameters

which describe the shape of the distribution or density function; these moments are most easily generated by characteristic functions. These topics are the subjects of Sections 3.6–3.8.

Because Gaussian random variables will form a basis of our system model, they are described in detail in Section 3.9. The initial model will involve linear operations on Gaussian inputs, so the results of such operations are discussed in Section 3.10. Finally, Section 3.11 solves the optimal estimation problem for cases in which a static linear Gaussian system model is an adequate description.

3.2 PROBABILITY AND RANDOM VARIABLES

Probability theory can be developed in an intuitive manner by describing probabilities of events of interest in terms of the *relative frequency of occurrence.* Using this concept, the probability of some event A, denoted as $P(A)$, can be generated as follows: if the event A is observed to occur $N(A)$ times in a total of N trials, then $P(A)$ is defined by

$$P(A) \triangleq \lim_{N \to \infty} \frac{N(A)}{N} \tag{3-1}$$

provided that this limit in fact exists. In other words, we conduct a number of experimental trials and observe the ratio of the number of times the event of interest occurs to the total number of trials. As we make more and more trials, if this ratio converges to some value, we call that value the probability of the event of interest.

Although this is a conceptually appealing basis for probability theory, it does not allow precise treatment of many problems and issues of direct importance to us. Modern probability theory is more rigorously based on an axiomatic definition of probability. This axiomatic definition must still be a valid mathematical model of empirically observed frequencies of occurrence, but it is meant to extract the essence of the ideas involved and to deal with them in a precise, rather than heuristic, manner.

To describe an experiment in precise terms, let Ω be the fundamental *sample space* containing all possible outcomes of the experiment conducted. Each single *elementary outcome* of the experiment is denoted as an ω; these ω's then are the elements of Ω: $\omega \in \Omega$. In other words, the sample space is just the collection of possible outcomes of the experiment, each of which being thought of as a point in that space Ω. Now let A be defined as a specific *event* of interest, a specific set of outcomes of the experiment. Thus, each such event A is a subset of Ω: $A \subset \Omega$. An event A is said to occur if the observed outcome ω is an element of A, if $\omega \in A$.

EXAMPLE 3.1 Consider two consecutive tosses of a fair coin. The sample space Ω is composed of four elements ω: if HT represents heads on the first throw and tails on the second, and

so forth, then the four possible elementary outcomes are HH, HT, TH, and TT. This is depicted schematically in Fig. 3.1. Ω is just the collection of those four outcomes.

Let us say we are interested in three events:

$$A_1 = \text{at least one tail was thrown}$$
$$A_2 = \text{exactly one tail was thrown}$$
$$A_3 = \text{exactly two tails were thrown}$$

Then A_1, A_2, and A_3 are subsets of Ω; each is a collection of points ω. These are also depicted in Fig. 3.1.

Now suppose we conduct one trial of the experiment, and we observe HT $= \omega_2$. Then, since this point is an element of sets A_1 and A_2, but not of A_3, we say that the events A_1 and A_2 occurred on that trial, but event A_3 did not occur. ■

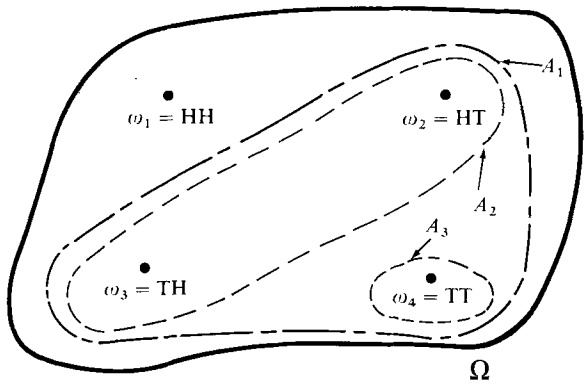

FIG. 3.1 Two tosses of a coin.

The sample space Ω can be discrete, with a finite or countably infinite number of elements, such as the space for the coin toss experiment described in the previous example. On the other hand, it could also be continuous, with an uncountable number of elements, such as the space appropriate to describe the continuous range of possible voltage values across a certain capacitor in a circuit.

So far, we have the structure of a sample space Ω, composed of elements ω_i, whose subsets are denoted as A_i. This is represented in Fig. 3.2.

Now we are going to restrict our attention to a certain class of sets A_i, a broad class called a σ-algebra, denoted as $\mathcal{F}$. In other words, the sets A_1, A_2, $A_3, \ldots$, which will be admissible for consideration will be elements of the class $\mathcal{F}$: $A_i \in \mathcal{F}$.

A *σ-algebra* $\mathcal{F}$ is a class of sets A_i, each of which is a subset of Ω ($A_i \subset \Omega$), such that if A_i is an element of $\mathcal{F}$ (i.e., if $A_i \in \mathcal{F}$), then:

(1) $A_i^* \in \mathcal{F}$, where A_i^* is the complement of A_i, $A_i^* = \Omega - A_i$,

(2) $\Omega \in \mathcal{F}$ [and then the empty set $\varnothing \in \mathcal{F}$ also, due to the preceding (1)],

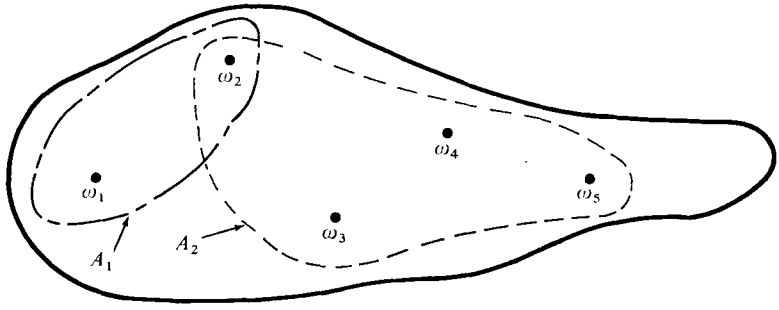

FIG. 3.2 Sample space Ω.

(3) if $A_1, A_2, \ldots \in \mathcal{F}$, then their union and intersection are also in $\mathcal{F}$:

$$\bigcup_i A_i \in \mathcal{F} \qquad \text{and} \qquad \bigcap_i A_i \in \mathcal{F}$$

where all possible finite and countably infinite unions and intersections are included.

Whereas Ω is a collection of points (elementary outcomes, ω), $\mathcal{F}$ is a collection of sets (events A_i), one of which is Ω itself.

For the purposes of our applications, we can let the sample space Ω be the set of points in n-dimensional Euclidean space R^n and let $\mathcal{F}$ be the class of sets generated by sets of the form (each of which is a subset of Ω):

$$A = \{\boldsymbol{\omega}: \boldsymbol{\omega} \leq \mathbf{a}, \boldsymbol{\omega} \in \Omega\} \tag{3-2}$$

and their complements, unions, and intersections. The notation in Eq. (3-2) requires some explanation. In other words, A is the set of $\boldsymbol{\omega}$'s that are elements of Ω (vectors in the n-dimensional Euclidean space, and thus, the boldfacing of ω to denote vector quantity; in the general case, ω will not be boldfaced); such that (:) $\boldsymbol{\omega} \leq \mathbf{a}$, where $\boldsymbol{\omega}$ and $\mathbf{a}$ are n-dimensional vectors and $\mathbf{a}$ is specified. Furthermore, $\boldsymbol{\omega} \leq \mathbf{a}$ is to be interpreted componentwise: $\boldsymbol{\omega} \leq \mathbf{a}$ means $\omega_1 \leq a_1$, $\omega_2 \leq a_2, \ldots, \omega_n \leq a_n$ for the n components ω_i and a_i of $\boldsymbol{\omega}$ and $\mathbf{a}$, respectively. This particular σ-algebra is of sufficient interest to have acquired its own name, and it is called a *Borel field*, denoted as $\mathcal{F}_B$. Taking complements, unions, and intersections of sets described by (3-2) leads to finite intervals (open, closed, half open) and point values along each of the n directions. Thus, a Borel field is composed of virtually all subsets of Euclidean n-space (R^n) that might be of interest in describing a probability problem associated with $\Omega = R^n$.

EXAMPLE 3.2 Consider generation of such sets of interest for $\Omega = R^1$, the real line. Let a_1 and a_2 be points on the real line, with $a_1 < a_2$. Then let

$$A_1 = \{\omega: \omega \leq a_1, \omega \in R^1\} = (-\infty, a_1]$$
$$A_2 = \{\omega: \omega \leq a_2, \omega \in R^1\} = (-\infty, a_2]$$

The complement of A_1, which is also a member of $\mathcal{F}_B$ by the definition of a σ-algebra, is

$$A_1^* = (-\infty, a_1]^* = (a_1, \infty)$$

Then the intersection of A_1^* and A_2 is

$$A_1^* \cap A_2 = (a_1, a_2]$$

Thus, we are able to generate any half-open interval, open on the left.

To generate points, we can look at a countably infinite intersection of half-open sets of the form

$$B_K = \{\omega : (b - \{1/K\}) < \omega \leq b\} = (b - \{1/K\}, b]$$

to generate

$$\bigcap_{K=1}^{\infty} B_K = \{\text{the point, } \omega = b\}$$

Note that we needed an infinite, not just finite, intersection to generate the point, and thus we needed to assure that such countable intersections yield sets in the σ-algebra when we first defined σ-algebra.

With a point b and a set $(b, c]$, we can generate the closed set $[b, c]$ by a simple union. Complementing and intersecting then yields open and half-open (open on the right) sets.

Thus, as claimed, the Borel field on the real line includes essentially all sets of possible interest. ■

Now define the *probability function* (or probability measure) $P(\cdot)$ to be a real scalar-valued function defined on $\mathcal{F}$ that assigns a value, $P(A)$, to each A which is a member of $\mathcal{F}$ ($A \in \mathcal{F}$) such that:

(1) $P(A) \geq 0$ for all $A \in \mathcal{F}$,

(2) $P(\Omega) = 1$,

(3) if $A_1, A_2, \ldots$ are elements of $\mathcal{F}$ and are disjoint, or mutually exclusive: i.e., if

$$A_i \cap A_j = \varnothing \qquad \text{for all} \quad i \neq j$$

then

$$P\left(\bigcup_{i=1}^{N} A_i\right) = \sum_{i=1}^{N} P(A_i)$$

for all finite and countably infinite N.

Such a definition for the function to determine the probability of the events $A \in \mathcal{F}$ does correspond to one's intuition of probability gained through the concept of relative frequency of occurrence. Each set of interest (i.e., each $A \in \mathcal{F}$) is assigned a probability value between 0 and 1 (P is a mapping from $\mathcal{F}$ into $[0, 1]$), and the probability of the sure event is one. Moreover, if A_1 is a subset of A_2, then the probability of set A_2 is at least as great as the probability of A_1, as expected. That this is a direct consequence of the axiomatic approach can be seen from Fig. 3.3. The set A_2 can be decomposed into two disjoint sets,

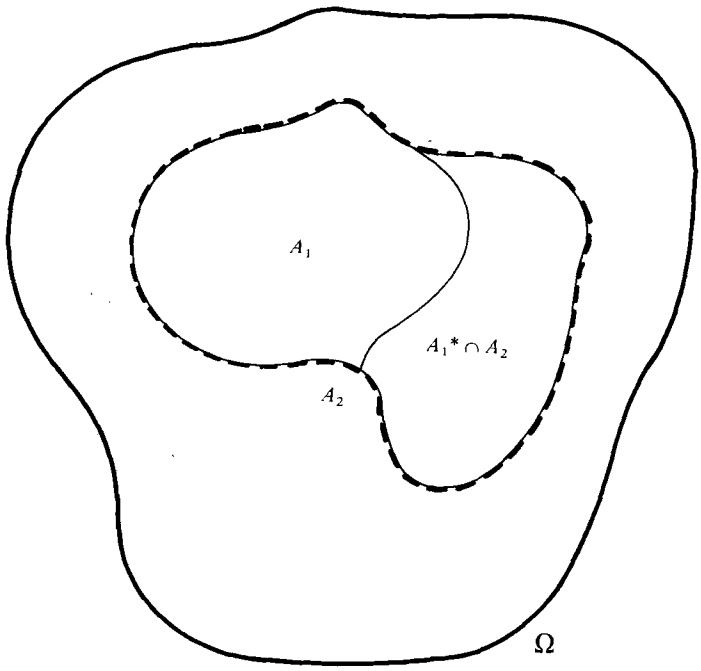

FIG. 3.3 $A_1 \subset A_2$ implies $P(A_1) \leq P(A_2)$.

A_1 and $A_1^* \cap A_2$. Then according to part (3) of the definition of a probability function, $P(A_2) = P(A_1) + P(A_1^* \cap A_2)$. From part (1), $P(A_1^* \cap A_2) \geq 0$, and so $P(A_2) \geq P(A_1)$ as desired.

Now we have what is called a *probability space*, defined by the triplet $(\Omega, \mathcal{F}, P)$ of the sample space, the underlying σ-algebra, and the probability function, all defined axiomatically as in the preceding. This entity serves as the basis of rigorously developed probability theory. Besides yielding results consistent with our intuitions about probability, this approach allows us to probe the essence of a problem and to determine whether or not it is posed properly. Subsequently, this rigor will also allow us to ensure that our definition of a random variable for a particular problem is in fact an appropriate choice for that application.

EXAMPLE 3.3 Let us consider the toss of a die to investigate a probability problem in the context of a rigorously defined probability space. Suppose that, for some reason, you are interested only in the occurrence of one of two events,

$$A_1 = \{\text{a 1 or a 2 was thrown}\} = \{1 \text{ or } 2\} \quad \text{and} \quad A_2 = \{\text{a 3 was thrown}\} = \{3\}$$

First of all, the sample space Ω is made up of the six possible outcomes: $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, and $\{6\}$.

One possible means of generating a σ-algebra would be to let $\mathcal{F}_1$ be composed of $\varnothing$, $\Omega = \{1, 2, 3, 4, 5, \text{ or } 6\}$, the six elementary outcomes (ω's), all possible unions of the outcomes two at a time, all possible unions three at a time, and so forth. A probability function would then have to assign a probability value to all such sets in $\mathcal{F}_1$. However, if one is only interested in A_1 and A_2 just defined, there is no need to be able to assign probabilities to such sets as {3 or 4 or 5}.

Another means of generating an appropriate σ-algebra, a "minimal" σ-algebra for this particular example, would be to let $\mathcal{F}_2$ be composed of $\varnothing$, Ω, A_1, A_2, and all possible complements, unions, and intersections thereof. Thus, $\mathcal{F}_2$ is made up of $\varnothing$, Ω, {1 or 2}, {3}, {1 or 2}* = {3 or 4 or 5 or 6}, {3}* = {1 or 2 or 4 or 5 or 6}, {1 or 2} $\cup$ {3} = {1 or 2 or 3}, {1 or 2 or 3}* = {4 or 5 or 6}. Conceptually, these are the only sets for which a probability must be defined in order to determine solutions to any probability questions posed in terms of events A_1 and A_2. Experiments could be conducted to assign probabilities to *only* these sets through the idea of relative frequency of occurrence, and the data generated would be complete.

Now let $P(A_1) = P(\{1 \text{ or } 2\}) = P_1$ and $P(A_2) = P(\{3\}) = P_2$. Then the probability function $P(\cdot)$ would assign probabilities to all of the elements of $\mathcal{F}$ as follows:

$$P(\varnothing) = 0 \qquad P(A_1^*) = P(\{3 \text{ or } 4 \text{ or } 5 \text{ or } 6\}) = 1 - P_1$$
$$P(\Omega) = 1 \qquad P(A_2^*) = P(\{1 \text{ or } 2 \text{ or } 4 \text{ or } 5 \text{ or } 6\}) = 1 - P_2$$
$$P(A_1) = P_1 \qquad P(A_1 \cup A_2) = P(\{1 \text{ or } 2 \text{ or } 3\}) = P_1 + P_2$$
$$P(A_2) = P_2 \qquad P(\{A_1 \cup A_2\}^*) = P(\{4 \text{ or } 5 \text{ or } 6\}) = 1 - P_1 - P_2$$

$P(A_1^*)$ is established by the fact that A_1 and A_1^* are disjoint sets whose union is Ω, so $P(A_1 \cup A_1^*) = 1 = P(A_1) + P(A_1^*) = P_1 + P(A_1^*)$; similarly for $P(A_2^*)$ and $P(\{A_1 \cup A_2\}^*)$. Since A_1 and A_2 are disjoint, $P(A_1 \cup A_2) = P(A_1) + P(A_2) = P_1 + P_2$. These are established by the axiomatic definitions, and would be verifiable by experimental observation (the theory is just abstractly modeling empirical results). ■

Once a probability space is properly defined for a given problem, the probability of all events of interest can be established, and theoretically we could be finished. The sample space Ω defines the possible outcomes of the experiment, $\mathcal{F}$ is the collection of events (sets) of interest, and P assigns a probability to every one of these events. However, we can deal with numerical representations of sets in a space more readily than with the abstract subsets themselves. Consequently, for quantitative analysis, we need a mapping from the sample space Ω to the real numbers. It is for this reason that we introduce the concept of a random variable.

A scalar *random variable* $\mathsf{x}(\cdot)$ is a real-valued point *function* which assigns a real scalar value to each point ω in Ω, denoted as $\mathsf{x}(\omega) = x$, such that every set $A \subset \Omega$ of the form

$$A = \{\omega : \mathsf{x}(\omega) \leq \xi\} \tag{3-3}$$

for ξ any value on the real line ($\xi \in R^1$), is an element of the σ-algebra $\mathcal{F}$ (i.e., $A \in \mathcal{F}$). The name "random variable" is perhaps unfortunate in that it does not seem to imply the fact that we are talking about a function, as opposed to values the function can assume. In fact, $\mathsf{x}(\cdot)$ is a function, or mapping, from Ω into R^1.

The notation used warrants discussion. Random variables will be set in sans serif type, $\mathsf{x}(\cdot)$ or simply x, to emphasize the fact that they are functions of

point values ω from the sample space Ω. The value that this function assumes for a particular ω, a *realization* of the random variable, will be the corresponding italicized letter. The corresponding Greek symbol will be used to denote a given vector or dummy variable (as, for integration), in the space of realizations. Thus, the notation $\{\omega: \mathsf{x}(\omega) \leq \xi\}$ is meant to read, "the set of ω in Ω such that the values assumed by the random variable function $\mathsf{x}(\cdot)$, for those ω as its argument, $\mathsf{x}(\omega) = x$, are less than or equal to the given number ξ on the real line."

A *vector random variable* or *random vector* $\mathbf{x}(\cdot)$ is just the generalization of the random variable concept to the vector case: a real-valued point function which assigns a real vector value to each point ω in Ω, denoted as $\mathbf{x}(\omega)$, such that every set A of the form

$$A = \{\omega: \mathbf{x}(\omega) \leq \boldsymbol{\xi}\} \tag{3-4}$$

for any $\boldsymbol{\xi} \in R^n$, is an element of $\mathcal{F}$. Although these definitions might at first seem contorted, there is good reason for their form. Scalar random variables are specifically mappings from Ω into R^1 such that inverse images of half-open intervals of the form $(-\infty, \xi]$ in R^1 are events in Ω that belong to $\mathcal{F}$. That is to say, they are events in Ω for which probabilities have been defined through the probability function P. Vector random variables are simply extensions of the same idea—mappings from Ω into R^n such that inverse images of sets of the form $\{\mathbf{x}(\omega) \in R^n: -\infty < \mathsf{x}_i(\omega) \leq \xi_i;\ i = 1, 2, \ldots, n\}$ are events in Ω to which probabilities have been ascribed. (From a measure theoretic point of view, this just says that random variables are measurable functions.)

EXAMPLE 3.4 Perhaps the best way to understand the concept of a random variable is to consider a function that is *not* a random variable for a given problem. Let Ω be the interval $(0, 10]$ on the real line, and suppose we are interested in distinguishing whether ω takes on a value in the interval $I_1 = (0, 5]$ or in $I_2 = (5, 10]$. The minimal σ-algebra $\mathcal{F}$ is made up of all possible complements, unions, and intersections of these two intervals, so that

$$\mathcal{F} = \{\varnothing,\ \Omega = (0, 10],\ I_1 = (0, 5],\ I_2 = (5, 10]\}$$

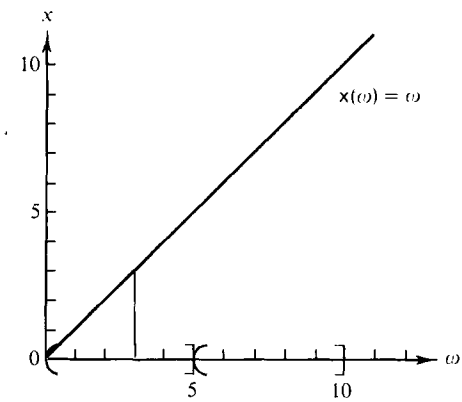

FIG. 3.4 Function that is not a random variable.

The value of ω can be anywhere along the line segment $(0, 10]$, and we just want to tell in which half of the segment it lies.

Now we want to establish an appropriate form for the function x to assume. Try defining $\mathsf{x}(\cdot)$ such that $\mathsf{x}(\omega) = \omega$, as in Fig. 3.4. This is not a suitable choice. Choose, for example, $\xi = 3$, as shown. Then the set A defined by

$$A = \{\omega : \mathsf{x}(\omega) \leq 3\} = (0, 3]$$

is *not* an element of the class $\mathscr{F}$. By definition, a random variable x *must* be defined such that all sets of the form

$$A = \{\omega : \mathsf{x}(\omega) \leq \xi\}$$

are in $\mathscr{F}$, for any choice of $\xi \in R^1$.

For this example, we must define x as assuming a constant value over $(0, 5]$ and a (different) constant over $(5, 10]$. One such random variable is shown in Fig. 3.5. Note for instance, that for this definition of x,

$$A_1 = \{\omega : \mathsf{x}(\omega) \leq 3\} = \varnothing; \qquad A_2 = \{\omega : \mathsf{x}(\omega) \leq 6\} = (0, 5]; \qquad A_3 = \{\omega : \mathsf{x}(\omega) \leq 20\} = (0, 10]$$

are all elements of $\mathscr{F}$. ■

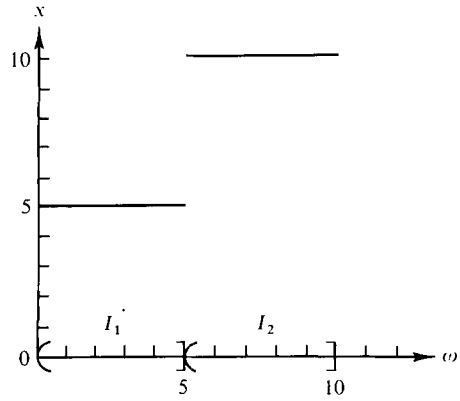

FIG. 3.5 Random variable definition. $\mathsf{x}(\omega) = 5$ if $\omega \in I_1$ and $\mathsf{x}(\omega) = 10$ if $\omega \in I_2$.

The sets I_1 and I_2 in the preceding example were irreducible elements of the σ-algebra $\mathscr{F}$, and a proper definition of a random variable required the function x to assume constant values over these sets. To generalize this concept, we call the set (event) $A \subset \Omega$ an "atom" of the σ-algebra $\mathscr{F}$ if $A \in \mathscr{F}$ and no subset of A is an element of $\mathscr{F}$ other than A itself and the null set $\varnothing$. A random variable can only assume a single value on an atom of the underlying σ-algebra $\mathscr{F}$.

Thus, we have the relationships depicted in Fig. 3.6. A random variable is a mapping from the fundamental sample space Ω into Euclidean n-space R^n. Each atom in Ω is mapped into a single vector in R^n. Conversely, the inverse image of sets in R^n of the form

$$A_i = \{\mathbf{x}(\omega) \in R^n : -\infty < \mathsf{x}_i(\omega) \leq \xi_i;\ i = 1, 2, \ldots, n\}$$

are events in Ω ($A_i \subset \Omega$) for which probabilities have been defined ($A_i \in \mathscr{F}$).

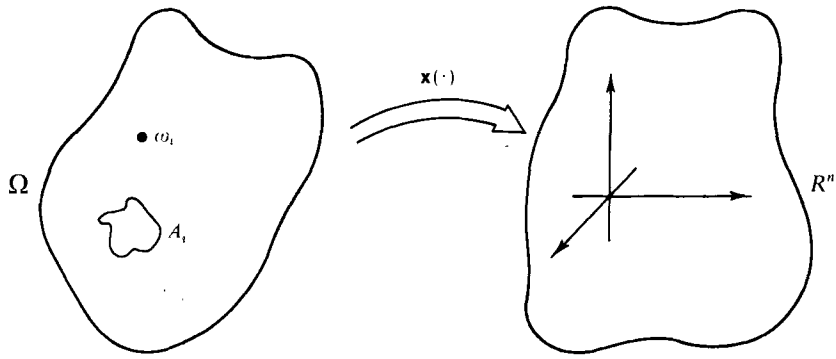

FIG. 3.6 The random variable mapping.

For the problems we will address in the sequel, the sample space Ω is R^n itself and the underlying σ-algebra is the Borel field $\mathcal{F}_B$ generated by sets of the form $A_i = \{\omega : \omega \leq \mathbf{a}, \omega \in \Omega\}$. An appropriate random variable definition for this case is simply the identity mapping suggested in Example 3.4:

$$\mathbf{x}(\omega) = \omega \tag{3-5}$$

Note that an atom in $\Omega = R^n$ is just a single point in the space (a single vector), and that the random variable just mentioned does map each such atom into a single vector in R^n. Thus, each realization $\mathbf{x}(\omega)$ is an n-dimensional vector whose components can take on any value in $(-\infty, \infty)$.

By the definition of a random variable $\mathbf{x}$, all sets of the form

$$A = \{\omega : \mathbf{x}(\omega) \leq \boldsymbol{\xi}\} = \{\omega : \mathsf{x}_1(\omega) \leq \xi_1, \mathsf{x}_2(\omega) \leq \xi_2, \ldots, \mathsf{x}_n(\omega) \leq \xi_n\}$$

have probabilities: probabilities are defined for them because $A \subset \Omega$ and $A \in \mathcal{F}$, and $P(\cdot)$ assigns probabilities to all such sets A. Therefore, the *probability distribution function* $F_{\mathbf{x}}(\cdot)$, a real scalar-valued function defined by

$$F_{\mathbf{x}}(\boldsymbol{\xi}) = P(\{\omega : \mathbf{x}(\omega) \leq \boldsymbol{\xi}\}) \tag{3-6a}$$

$$= \text{“}P(\mathbf{x} \leq \boldsymbol{\xi})\text{”} \tag{3-6b}$$

$$= \text{“}P(\mathsf{x}_1 \leq \xi_1, \mathsf{x}_2 \leq \xi_2, \ldots, \mathsf{x}_n \leq \xi_n)\text{”} \tag{3-6c}$$

always exists. We have defined the various sets and functions to this point so as to assure that such a function exists. The quotation marks in (3-6) are meant to emphasize that such notation, very typical in probability theory literature, should be interpreted in terms of the probability of a set of ω's in the original sample space Ω. Sometimes the notation $F(\boldsymbol{\xi})$ is used rather than $F_{\mathbf{x}}(\boldsymbol{\xi})$, if the random variable concerned is obvious from context. Moreover, since

$$F_{\mathbf{x}}(\boldsymbol{\xi}) = F_{\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n}(\xi_1, \xi_2, \ldots, \xi_n) \tag{3-7}$$

this is sometimes called the joint probability distribution function of $\mathsf{x}_1, \mathsf{x}_2, \ldots,$ and x_n.

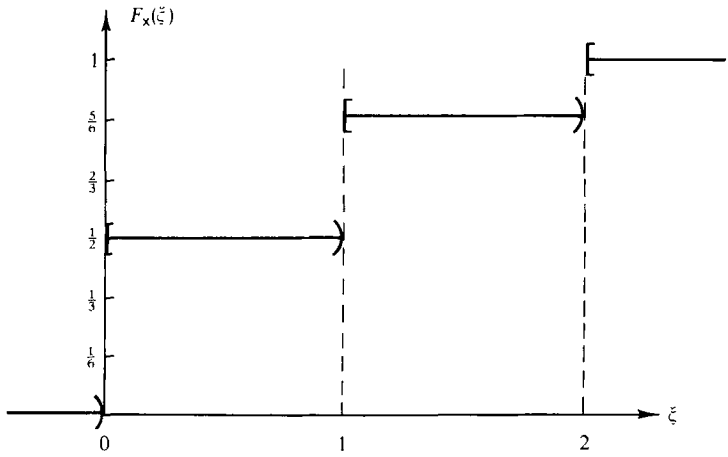

FIG. 3.7 Probability distribution function.

EXAMPLE 3.5 Consider the die-toss experiment introduced in Example 3.3, in which we were interested only in the sets $A_1 = \{1 \text{ or } 2\}$ and $A_2 = \{3\}$. Define a random variable x through

$$\mathsf{x}(\omega) = \begin{cases} 0 & \text{if } \omega \notin A_1 \text{ or } A_2 \\ 1 & \text{if } \omega \in A_1 \\ 2 & \text{if } \omega \in A_2 \end{cases}$$

As in Example 3.3, let $P(A_1) = P_1$ and $P(A_2) = P_2$ ($\frac{1}{3}$ and $\frac{1}{6}$, respectively, for a fair die).

Now we will establish the probability distribution function

$$F(\xi) = P(\{\omega : \mathsf{x}(\omega) \leq \xi\})$$

For $\xi < 0$, $P(\{\omega : \mathsf{x}(\omega) \leq \xi\}) = P(\varnothing) = 0$.

For $0 \leq \xi < 1$, $P(\{\omega : \mathsf{x}(\omega) \leq \xi\}) = P(\{A_1 \cup A_2\}^*) = 1 - P_1 - P_2 = \frac{1}{2}$.

For $1 \leq \xi < 2$, $P(\{\omega : \mathsf{x}(\omega) \leq \xi\}) = P(A_2^*) = 1 - P_2 = \frac{5}{6}$.

For $2 \leq \xi < \infty$, $P(\{\omega : \mathsf{x}(\omega) \leq \xi\}) = P(\Omega) = 1$.

Plotting $F_x(\xi)$ versus ξ yields the graphical depiction of the probability distribution function in Fig. 3.7. ■

Figure 3.8 summarizes the concepts that have been discussed. We started with an abstract sample space Ω, composed of elements (points) ω that were the elementary outcomes of an experiment. There were also certain subsets A of Ω ($A \subset \Omega$) of interest, called events, and specifically these sets were from a class $\mathcal{F}$ ($A \in \mathcal{F}$) called a σ-algebra. For any $A \in \mathcal{F}$, we could evaluate the *set function* $P(\cdot)$, a mapping from $\mathcal{F}$ into $[0, 1]$, to generate probabilities as $P(A)$. The triplet $(\Omega, \mathcal{F}, P)$ then defined what was termed a probability space.

We also defined a *point function* $\mathbf{x}(\cdot)$ called a random variable, a mapping from Ω into R^n, which could be evaluated for each $\omega \in \Omega$ to yield realizations $\mathbf{x}(\omega)$. The probabilities established as $P(A)$ and the realizations $\mathbf{x}(\omega)$ of the random variable $\mathbf{x}$ are then related by the probability distribution function $F_{\mathbf{x}}(\cdot)$, a mapping from R^n into $[0, 1]$, that yields $F_{\mathbf{x}}(\boldsymbol{\xi})$ as the probability of the set of $\omega \in \Omega$ such that $\mathbf{x}(\omega) \leq \boldsymbol{\xi}$.

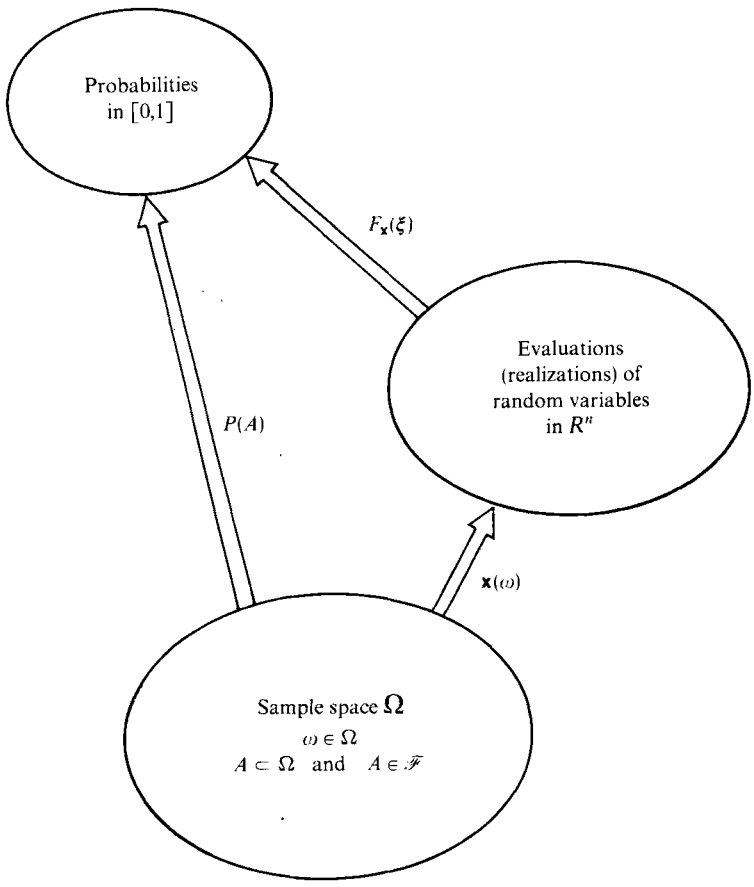

FIG. 3.8 Probability and random variables.

3.3 PROBABILITY DISTRIBUTIONS AND DENSITIES

As discussed in the previous section, the probability distribution function is a basic entity associated with any random variable that allows us to generate probabilities of sets of interest. We are assured of its existence. On the other hand, we are not assured of the existence of its derivative everywhere, but if it does exist, it is often easier to use and more revealing in terms of graphical interpretations.

Given a vector random variable $\mathbf{x}$,

$$\mathbf{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{3-8}$$

the probability distribution function $F_{\mathbf{x}}$ can be evaluated as a scalar function of the dummy vector $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_n]^T$:

$$F_{\mathbf{x}}(\boldsymbol{\xi}) \triangleq F_{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n}(\xi_1, \xi_2, \ldots, \xi_n) \tag{3-9a}$$

$$\triangleq P(\{\omega : \mathbf{x}_1(\omega) \leq \xi_1, \mathbf{x}_2(\omega) \leq \xi_2, \ldots, \mathbf{x}_n(\omega) \leq \xi_n\}) \tag{3-9b}$$

Note again that we specifically avoid the notation $F_{\mathbf{x}}(\mathbf{x})$, as is often used, to prevent giving the impression that $F_{\mathbf{x}}$ is in any way a function of particular realizations $\mathbf{x}$ of $\mathbf{x}$. As can be seen from Eq. (3-9), $F_{\mathbf{x}}$ is a monotonic nondecreasing function of any component ξ_i of the vector $\boldsymbol{\xi}$: for instance, the probability of the set of ω such that $\mathbf{x}_i(\omega) \leq 2$ must be at least as large as the probability of the set of ω such that $\mathbf{x}_i(\omega) \leq 1$.

Other properties of this function that become apparent from its definition include:

$$F_{\mathbf{x}}(\infty, \infty, \ldots, \infty) = P(\{\omega : \mathbf{x}_1(\omega) \leq \infty, \ldots, \mathbf{x}_n(\omega) \leq \infty\}) = 1 \tag{3-10}$$

$$F_{\mathbf{x}}(\xi_1, \ldots, -\infty, \ldots, \xi_n) = P\{\omega : \ldots, \mathbf{x}_i(\omega) \leq -\infty, \ldots\}) = 0 \tag{3-11}$$

If all of its arguments are ∞, the value $F_{\mathbf{x}}$ assumes is one; if any single argument is $-\infty$, it takes on the value zero (these statements are more properly expressed in the limit as certain arguments tend to $+\infty$ or $-\infty$). If we are interested only in probabilities concerning the first k of the n random variables, and $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$ can take on any values, then:

$$\begin{aligned} F_{\mathbf{x}_1, \ldots, \mathbf{x}_k}(\xi_1, \ldots, \xi_k) &= P(\{\omega : \mathbf{x}_1(\omega) \leq \xi_1, \ldots, \mathbf{x}_k(\omega) \leq \xi_k\}) \\ &= P(\{\omega : \mathbf{x}_1(\omega) \leq \xi_1, \ldots, \mathbf{x}_k(\omega) \leq \xi_k, \\ &\quad \mathbf{x}_{k+1}(\omega) \leq \infty, \ldots, \mathbf{x}_n(\omega) \leq \infty\}) \\ &= F_{\mathbf{x}_1, \ldots, \mathbf{x}_n}(\xi_1, \ldots, \xi_k, \infty, \ldots, \infty) \end{aligned} \tag{3-12}$$

Equation (3-12) embodies the concept of a "*marginal*" *probability distribution* of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Note that the first k components were chosen only so Eq. (3-12) could be written readily; any argument of ∞ in $F_{\mathbf{x}}(\boldsymbol{\xi})$ yields the corresponding result [the variables can be reordered so there is no loss of generality in (3-12)].

The probability distribution function can be used to generate probabilities of other sets of interest as well. For instance, in the scalar case, the probability of sets of ω such that $\mathbf{x}$ assumes a value in a given half-open interval $(\xi_1, \xi_2]$ can be readily established. The set $\{\omega : \mathbf{x}(\omega) \leq \xi_2\}$ can be decomposed into the union of two disjoint sets:

$$\{\omega : \mathbf{x}(\omega) \leq \xi_2\} = \{\omega : \mathbf{x}(\omega) \leq \xi_1\} \cup \{\omega : \mathbf{x}(\omega) \in (\xi_1, \xi_2]\}$$

Because the sets on the right are disjoint, we have

$$P(\{\omega : \mathbf{x}(\omega) \leq \xi_2\}) = P(\{\omega : \mathbf{x}(\omega) \leq \xi_1\}) + P(\{\omega : \mathbf{x}(\omega) \in (\xi_1, \xi_2]\})$$

so we can write

$$P(\{\omega:\mathsf{x}(\omega)\in(\xi_1,\xi_2]\}) = F_{\mathsf{x}}(\xi_2) - F_{\mathsf{x}}(\xi_1) \tag{3-13}$$

To generate probabilities of open or closed sets, we need to evaluate probabilities that x assumes a single value. If the distribution function is discontinuous at some ξ_0, then there is a finite probability that $\mathsf{x}(\cdot)$ assumes that value. In (3-13), let $\xi_2 = \xi_0$ and $\xi_1 = \xi_0 - \varepsilon$, and take the limit as $\varepsilon \to 0$ to yield

$$P(\{\omega:\mathsf{x}(\omega) = \xi_0\}) = F_{\mathsf{x}}(\xi_0) - F_{\mathsf{x}}(\xi_0^-) \tag{3-14}$$

i.e., the probability is equal to the magnitude of the jump discontinuity. We have observed such discontinuities in Example 3.5. Thus, for instance, since we have disjoint sets,

$$\begin{aligned} P(\{\omega:\mathsf{x}(\omega)\in[\xi_1,\xi_2]\}) &= P(\{\omega:\mathsf{x}(\omega)=\xi_1\}) + P(\{\omega:\mathsf{x}(\omega)\in(\xi_1,\xi_2]\}) \\ &= [F_{\mathsf{x}}(\xi_1) - F_{\mathsf{x}}(\xi_1^-)] + [F_{\mathsf{x}}(\xi_2) - F_{\mathsf{x}}(\xi_1)] \\ &= F_{\mathsf{x}}(\xi_2) - F_{\mathsf{x}}(\xi_1^-) \end{aligned} \tag{3-15}$$

We will discuss the generation of probabilities for general sets of interest after we introduce the concept of a probability density function.

If a scalar-valued function $f_{\mathbf{x}}(\cdot)$ exists such that

$$F_{\mathbf{x}}(\xi_1,\xi_2,\ldots,\xi_n) = \int_{-\infty}^{\xi_1}\int_{-\infty}^{\xi_2}\cdots\int_{-\infty}^{\xi_n} f_{\mathbf{x}}(\rho_1,\rho_2,\ldots,\rho_n)\,d\rho_1\,d\rho_2\cdots d\rho_n \tag{3-16a}$$

or, in a simpler notation to represent the same expression,

$$F_{\mathbf{x}}(\boldsymbol{\xi}) = \int_{-\infty}^{\boldsymbol{\xi}} f_{\mathbf{x}}(\boldsymbol{\rho})\,d\boldsymbol{\rho} \tag{3-16b}$$

holds for all values of $\boldsymbol{\xi} = [\xi_1,\xi_2,\ldots,\xi_n]^{\mathrm{T}}$, then this function $f_{\mathbf{x}}$ is the *probability density function* of $\mathbf{x}$. Unlike the probability distribution function, we are not always assured of the existence of $f_{\mathbf{x}}$. If $F_{\mathbf{x}}$ is absolutely continuous, then the density function does exist (absolute continuity can be defined rigorously through measure theory, but basically $F_{\mathbf{x}}$ is absolutely continuous if the number of points where it is not differentiable is countable). If such a density function exists, then $\mathbf{x}$ is termed a *continuous random variable.*

By the fundamental theorem of calculus, we can use (3-16) to deduce

$$f_{\mathbf{x}}(\boldsymbol{\xi}) = \frac{\partial^n}{\partial\xi_1\,\partial\xi_2\cdots\partial\xi_n} F_{\mathbf{x}}(\boldsymbol{\xi}) \tag{3-17}$$

This relationship and (3-16), combined with properties of $F_{\mathbf{x}}$, yield some properties of $f_{\mathbf{x}}$. Since $F_{\mathbf{x}}$ is monotonic nondecreasing,

$$f_{\mathbf{x}}(\boldsymbol{\xi}) \geq 0 \qquad \text{for all} \quad \boldsymbol{\xi} \tag{3-18}$$

In view of (3-10), it is a property of a density function that

$$\int_{-\infty}^{\infty} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi} = 1 \tag{3-19}$$

If we are interested only in the first k of the n components of $\mathbf{x}$, and $\mathsf{x}_{k+1}, \ldots, \mathsf{x}_n$ can take on any values, then we can establish the *marginal density function* by integrating out the dependence upon the last $(k - n)$ components:

$$f_{\mathsf{x}_1,\ldots,\mathsf{x}_k}(\xi_1, \ldots, \xi_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathsf{x}_1,\ldots,\mathsf{x}_n}(\xi_1, \ldots, \xi_k, \ldots, \xi_n)\, d\xi_{k+1} \cdots d\xi_n \tag{3-20}$$

as can be seen from (3-12). With F_{x} continuous, (3-14) yields

$$P(\{\omega : \mathsf{x}(\omega) = \xi_0\}) = 0 \tag{3-21}$$

so that, using (3-13) and (3-15),

$$P(\{\omega : \mathsf{x}(\omega) \in (\xi_1, \xi_2] \text{ or } [\xi_1, \xi_2]\}) = F_{\mathsf{x}}(\xi_2) - F_{\mathsf{x}}(\xi_1) = \int_{\xi_1}^{\xi_2} f_{\mathsf{x}}(\rho)\, d\rho \tag{3-22}$$

with extensions to the vector case.

EXAMPLE 3.6 Two forms of random variables useful for modeling empirically observed phenomena are the uniformly and Gaussian (or normally) distributed random variables. In the scalar case, their probability distribution and density functions can be plotted as in Fig. 3.9.

The uniformly distributed random variable models a situation in which a quantity of interest can take on any value in a specified range (limited by physical considerations as gimbal stops on a servo motor, by definition of units as angular orientation being described in the range $[0, 2\pi)$, etc.) and in which there is no reason to believe certain ranges of values to be more probable than others. The Gaussian, or normal, random variable serves as a good model for many observed phenomena and will be discussed at length in Section 3.9.

Note that, analogous to a mass density function, the probability density function indicates where the probability (mass) is concentrated. It is partly this graphic portrayal of probable ranges of values that makes the density function more attractive to use than the distribution function. ■

Now we want to obtain the probability of general sets of interest associated with vector random variables. In the scalar case, we can see from (3-22) that the probability of the set of ω such that $\mathsf{x}(\omega)$ lies in the infinitesimal interval from ξ_1 to $(\xi_1 + d\xi_1)$ is just

$$P(\{\omega : \mathsf{x}(\omega) \in [\xi_1, \xi_1 + d\xi_1]\}) = f_{\mathsf{x}}(\xi_1)\, d\xi_1 \tag{3-23a}$$

This generalizes to the n-dimensional case as

$$P(\{\omega : \mathsf{x}_i(\omega) \in [\xi_i, \xi_i + d\xi_i];\ i = 1, 2, \ldots, n\}) = f_{\mathbf{x}}(\xi_1, \ldots, \xi_n)\, d\xi_1 \cdots d\xi_n \tag{3-23b}$$

Thus, the probability that the random variable takes on a value within an infinitesimal hypercube is just the *probability (mass) density* evaluated at the location of the hypercube (it is constant over the *infinitesimal* volume) multiplied by the *volume* of the hypercube, $[d\xi_1 \cdots d\xi_n]$. Then any set A in R^n can be generated from such hypercubes, and so the probability that $\mathbf{x}(\omega) = \mathbf{x}$ lies in

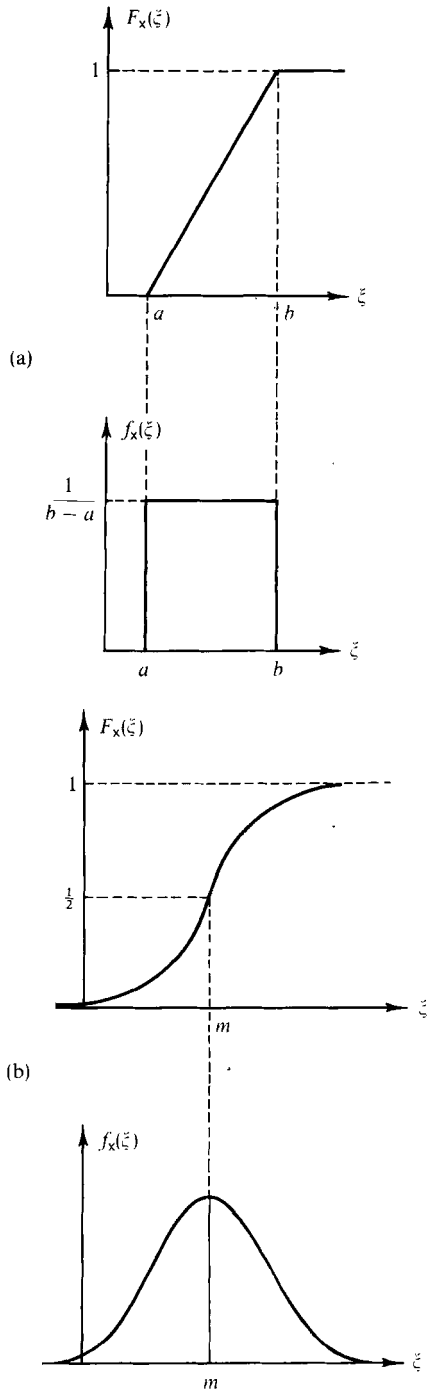

FIG. 3.9 (a) Uniform and (b) Gaussian (normal) distributions and densities.

the set A is

$$P(\{\omega:\mathbf{x}(\omega)\in A\}) = \int_A f_{\mathbf{x}}(\boldsymbol{\xi})\,d\boldsymbol{\xi} \tag{3-24a}$$

$$= \int_A \cdots \int f_{\mathbf{x}}(\xi_1,\ldots,\xi_n)\,d\xi_1\cdots d\xi_n \tag{3-24b}$$

In this manner, the probability associated with sets of interest can be generated through ordinary (Riemann) integration with the probability density function.

To think of this geometrically, consider Fig. 3.10. In case (a), x is scalar and $f_x(\xi)$ is a simple curve plotted against ξ. Sets A of interest are intervals along the abscissa (possibly disjoint), and $P(\{\omega:x(\omega)\in A\})$ can be determined as the area under the curve delimited by A. If $\mathbf{x}$ is two dimensional, as in case (b), $f_{\mathbf{x}}(\xi_1,\xi_2)$ is a surface over the ξ_1–ξ_2 plane; sets A are areas in the ξ_1–ξ_2 plane, and $P(\{\omega:\mathbf{x}(\omega)\in A\})$ can be calculated as the volume under the surface $f_{\mathbf{x}}(\xi_1,\xi_2)$ with A as a cross section.

If we want to calculate the probability of general sets without use of the density function (as, for cases in which $F_{\mathbf{x}}$ has discontinuities so $f_{\mathbf{x}}$ cannot be

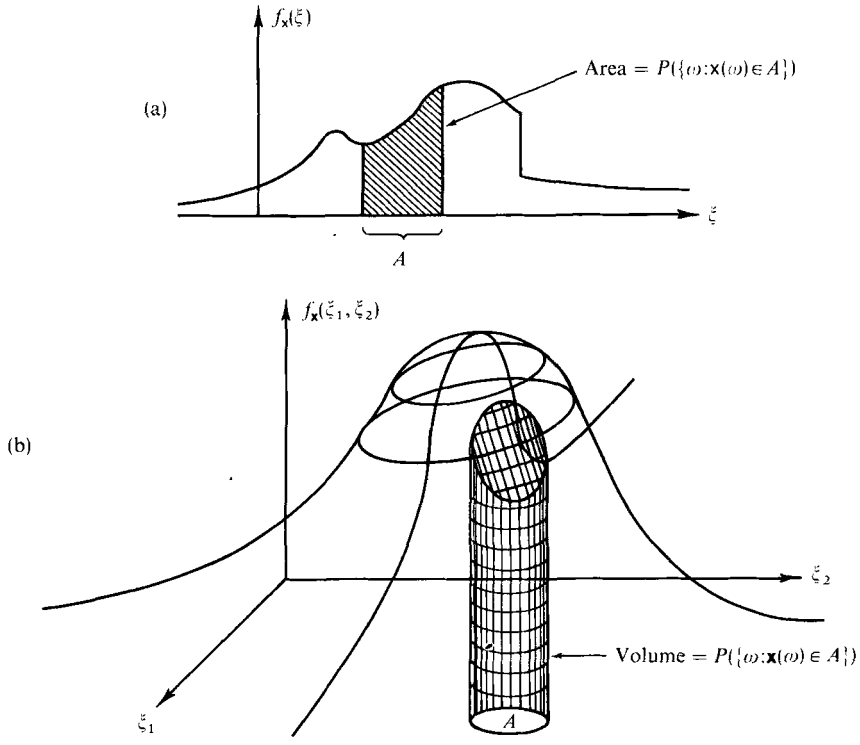

FIG. 3.10 Probability of sets. (a) Scalar-valued x. (b) Two-dimensional vector-valued $\mathbf{x}$.

defined everywhere), the preceding ideas can be extended through

$$P(\{\omega:\mathbf{x}(\omega) \in A\}) = \int_A dF_{\mathbf{x}}(\xi) \tag{3-25}$$

Meaning is given to this expression through measure theory: we need to define this Lebesgue–Stieltjes integral of $F_{\mathbf{x}}$ over A. If $f_{\mathbf{x}}$ exists, then (3-24) and (3-25) yield the same result, with (3-24) being more attractive because it is in terms of ordinary Riemann integration. For our applications, we will be able to assume the existence of $f_{\mathbf{x}}$. We will therefore be able to avoid measure theory considerations, but such an extension *can* be made.

Let us reflect on what has been accomplished through the random variable mapping $\mathbf{x}$. Recall Fig. 3.6: $\mathbf{x}$ maps Ω into R^n such that each atom (each irreducible set) in Ω maps into a single vector in R^n. Thus, the sets of interest in R^n will be elements of the Borel field $\mathscr{F}_B$ associated with R^n. For all sets $A \subset R^n$ and $A \in \mathscr{F}_B$, we can define probabilities through

$$P_x(A) = \int_A dF_{\mathbf{x}}(\xi) \tag{3-26}$$

where $P_x(\cdot)$ is the probability function (Borel measure) associated with R^n, or, if the probability density function exists, as

$$P_x'(A) = \int_A f_{\mathbf{x}}(\xi)\, d\xi = P_x(A) \tag{3-27}$$

Equation (3-24) relates that $P_x(A \subset R^n)$ defined in (3-26) is equal to $P(\{\omega:\mathbf{x}(\omega) \in A\} \subset \Omega)$ for all sets A of interest. We now have a *new probability space*, $(R^n, \mathscr{F}_B, P_x)$, generated by the mapping $\mathbf{x}$ from the original probability space:

$$(\Omega, \mathscr{F}, P) \xrightarrow{\mathbf{x}(\cdot)} (R^n, \mathscr{F}_B, P_x) \tag{3-28}$$

Quite often, one can describe a problem conveniently in terms of the probability space $(R^n, \mathscr{F}_B, P_x)$, neglecting the original probability space. For our applications, $\mathbf{x}(\cdot)$ is the identity mapping, so the issue is rather academic. However, there are many problem areas in which recollection of the fundamental probability space and associated sets of interest yields clear insights into subtle and troublesome considerations.

3.4 CONDITIONAL PROBABILITY AND DENSITIES

Suppose we have two random variables $\mathbf{x}$ and $\mathbf{y}$ mapping from a sample space Ω into R^n and R^m, respectively. Further suppose that $\mathbf{x}$ can assume only discrete values $\mathbf{x}_i$ and similarly $\mathbf{y}$ can take only discrete values $\mathbf{y}_j$, with i and j integers (a finite or countably infinite number). If we knew that $\mathbf{y}$ has assumed a particular realization $\mathbf{y}_j$, that knowledge would, in general, affect the determination of the probability that $\mathbf{x}(\omega) = \mathbf{x}_i$ for a given value $\mathbf{x}_i$. Figure 3.11

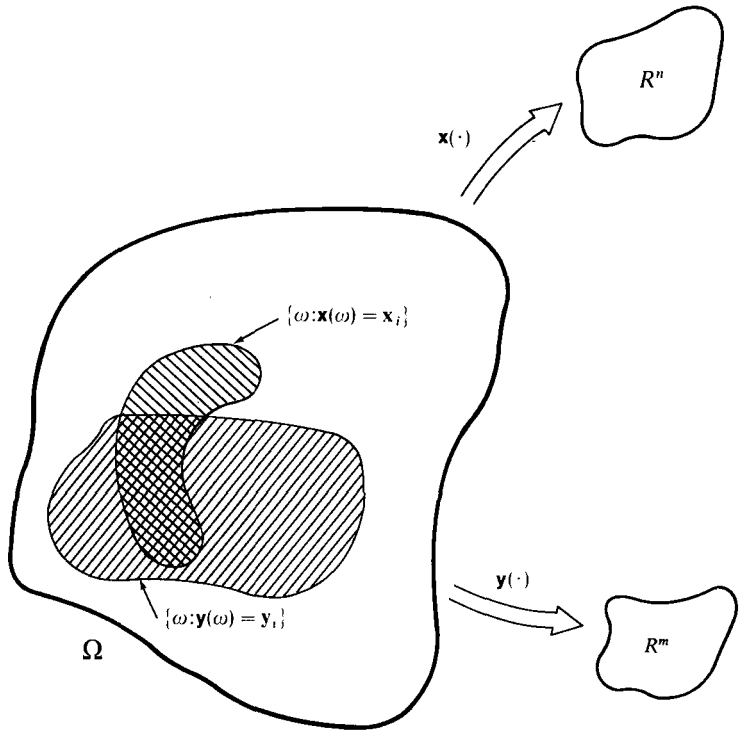

FIG. 3.11 Conditional probability via sets in Ω. The crosshatching is $\{\omega:\mathbf{x}(\omega) = \mathbf{x}_i$ and $\mathbf{y}(\omega) = \mathbf{y}_j\}$.

depicts the various sets of interest in the sample space Ω, namely $\{\omega:\mathbf{x}(\omega) = \mathbf{x}_i\}$ and $\{\omega:\mathbf{y}(\omega) = \mathbf{y}_j\}$. If we did not know the value $\mathbf{y}$ assumes, then we would simply evaluate $P(\{\omega:\mathbf{x}(\omega) = \mathbf{x}_i\})$ using the probability measure defined for sets A, $A \subset \Omega$ and $A \in \mathcal{F}$. However, if we *know* that $\mathbf{y}(\omega) = \mathbf{y}_j$, we can restrict our attention to $\{\omega:\mathbf{y}(\omega) = \mathbf{y}_j\} \subset \Omega$ instead of considering all Ω. *Within that set* of ω, we want to know the probability of the set of ω such that $\mathbf{x}(\omega) = \mathbf{x}_i$ as well: the probability of $\{\omega:\mathbf{x}(\omega) = \mathbf{x}_j$ and $\mathbf{y}(\omega) = \mathbf{y}_j\}$, the cross hatched set in Fig. 3.11, relative to the set $\{\omega:\mathbf{y}(\omega) = \mathbf{y}_j\}$. Thus, the conditional probability that $\mathbf{x}(\omega) = \mathbf{x}_i$, conditioned on the fact that $\mathbf{y}(\omega) = \mathbf{y}_j$, can be defined as

$$P(\mathbf{x}(\omega) = \mathbf{x}_i \,|\, \mathbf{y}(\omega) = \mathbf{y}_j) = \frac{P(\mathbf{x}(\omega) = \mathbf{x}_i \text{ and } \mathbf{y}(\omega) = \mathbf{y}_j)}{P(\mathbf{y}(\omega) = \mathbf{y}_j)} \tag{3-29a}$$

$$= \frac{P(\{\omega:\mathbf{x}(\omega) = \mathbf{x}_i \text{ and } \mathbf{y}(\omega) = \mathbf{y}_j\})}{P(\{\omega:\mathbf{y}(\omega) = \mathbf{y}_j\})} \tag{3-29b}$$

Note that this conditional probability need only be defined for sets $A \subset \{\omega : \mathbf{y}(\omega) = \mathbf{y}_j\} \subset \Omega$, $A \in \mathscr{F}' \subset \mathscr{F}$ ($\mathscr{F}'$ can be a "coarser" σ-algebra than $\mathscr{F}$, consisting of fewer elements).

To evaluate such a conditional probability numerically, first one could calculate the numerator in (3-29) as the ratio of the number of trials in which both events occur to the total number of trials. Then the denominator could be generated as the ratio of trials in which $\mathbf{y}(\omega) = \mathbf{y}_j$ to the total number of trials. Note also that if we summed over all $\mathbf{x}_i$'s, we would obtain a probability of one for any given $\mathbf{y}_j$.

However, this definition of conditional probability is valid only if $P(\mathbf{y}(\omega) = \mathbf{y}_j) > 0$. In our applications, we will be considering continuous random variables $\mathbf{y}$, for which $P(\mathbf{y}(\omega) = \mathbf{y}_j) = 0$, so the previous definition breaks down. Measure theory can be used to develop the concept of conditional probabilities rigorously and in more generality than we will require (we will assume the existence of appropriate densities). More will be said about this measure theoretic approach in Section 3.7, once the idea of expectation has been introduced. For now, we will develop the concept of a conditional density function, which will be of basic importance for estimation problems addressed in the sequel.

Let us first provide an interpretation of $f_{\mathbf{x}|\mathbf{y}}(\xi \,|\, \mathbf{y}_0)$, the conditional density of $\mathbf{x}$ as a function of ξ, conditioned on knowledge that the random variable $\mathbf{y}$ has assumed the realization $\mathbf{y}_0 : \mathbf{y}(\omega) = \mathbf{y}_0$. Let $\mathbf{x}$ map Ω into R^n and $\mathbf{y}$ map Ω into R^m, and let $A \subset R^n$ and $B \subset R^m$ be point sets of interest in the corresponding spaces, as in Fig. 3.12a. The conditional probability that $\mathbf{x}(\omega)$ lies in A, conditioned on the fact that $\mathbf{y}(\omega) \in B$, is

$$P(\mathbf{x}(\omega) \in A \,|\, \mathbf{y}(\omega) \in B) = \frac{P(\mathbf{x}(\omega) \in A \text{ and } \mathbf{y}(\omega) \in B)}{P(\mathbf{y}(\omega) \in B)} \tag{3-30}$$

provided that $P(\mathbf{y}(\omega) \in B)$ is nonzero. The probabilities on the right hand side of (3-30) can be evaluated using the probability function P associated with $(\Omega, \mathscr{F}, P)$, or with the functions P'_{xy} and P_y' associated with $(R^{nm}, \mathscr{F}_B, P'_{xy})$ and $(R^m, \mathscr{F}_B, P_y')$ as described in (3-27). Thus, letting $f_{\mathbf{x},\mathbf{y}}(\cdot,\cdot)$ denote the joint probability density of $\mathbf{x}$ and $\mathbf{y}$,

$$P(\mathbf{x}(\omega) \in A \,|\, \mathbf{y}(\omega) \in B) = \frac{\int_A \left[\int_B f_{\mathbf{x},\mathbf{y}}(\xi, \gamma)\, d\gamma\right] d\xi}{\int_B f_{\mathbf{y}}(\rho)\, d\rho} \tag{3-31a}$$

$$= \int_A \frac{\int_B f_{\mathbf{x},\mathbf{y}}(\xi, \gamma)\, d\gamma}{\int_B f_{\mathbf{y}}(\rho)\, d\rho}\, d\xi \tag{3-31b}$$

From (3-31b) it can be seen that the conditional density for $\mathbf{x}$, conditioned on the fact that $\mathbf{y}(\omega) \in B$, would be

$$f_{\mathbf{x}}(\xi \,|\, \mathbf{y}(\omega) \in B) = \frac{\int_B f_{\mathbf{x},\mathbf{y}}(\xi, \gamma)\, d\gamma}{\int_B f_{\mathbf{y}}(\rho)\, d\rho} \tag{3-32}$$

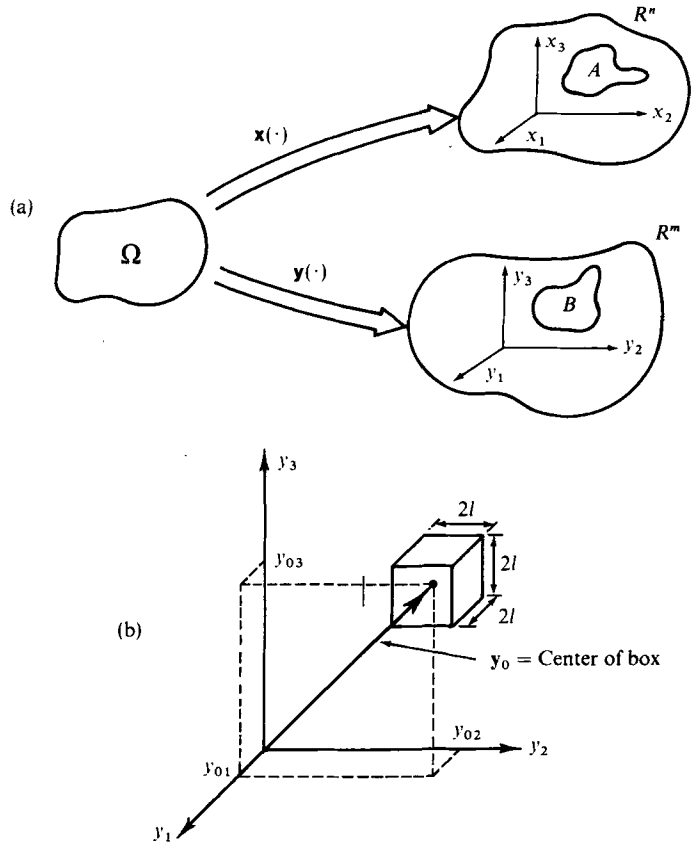

FIG. 3.12 (a) Sets of interest in R^n and R^m. (b) A particular set $B \subset R^m$. y_0 is center of box.

Now consider a particular set $B \subset R^m$, namely a hypercube centered at $\mathbf{y}_0$ of dimension $2l$ on each side, as shown in Fig. 3.12b:

$$B = \{\mathbf{y} \in R^m : |y_1 - y_{01}| \leq l, |y_2 - y_{02}| \leq l, \ldots, |y_m - y_{0m}| \leq l\} \quad (3\text{-}33)$$

We can write the density functions $f_{\mathbf{x},\mathbf{y}}(\xi, \gamma)$ and $f_{\mathbf{y}}(\rho)$ in terms of their evaluations at the given value of $\mathbf{y}_0$ as

$$f_{\mathbf{x},\mathbf{y}}(\xi, \gamma) = f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{y}_0) + \delta f_{\mathbf{x},\mathbf{y}}(\xi, \gamma - \mathbf{y}_0) \quad (3\text{-}34a)$$

$$f_{\mathbf{y}}(\rho) = f_{\mathbf{y}}(\mathbf{y}_0) + \delta f_{\mathbf{y}}(\rho - \mathbf{y}_0) \quad (3\text{-}34b)$$

Thus, (3-32) can be written as

$$f_{\mathbf{x}}(\xi \,|\, \mathbf{y}(\omega) \in B) = \frac{\int_B [f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{y}_0) + \delta f_{\mathbf{x},\mathbf{y}}(\xi, \gamma - \mathbf{y}_0)]\, d\gamma}{\int_B [f_{\mathbf{y}}(\mathbf{y}_0) + \delta f_{\mathbf{y}}(\rho - \mathbf{y}_0)]\, d\rho} \quad (3\text{-}35)$$

Now let V_B be the "volume" of the hypercube in m dimensions, $(2l)^m$, to write

$$f_{\mathbf{x}}(\xi \mid \mathbf{y}(\omega) \in B) = \frac{V_B f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{y}_0) + \int_B \delta f_{\mathbf{x},\mathbf{y}}(\xi, \gamma - \mathbf{y}_0)\, d\gamma}{V_B f_{\mathbf{y}}(\mathbf{y}_0) + \int_B \delta f_{\mathbf{y}}(\rho - \mathbf{y}_0)\, d\rho}$$

$$= \frac{f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{y}_0) + (1/V_B) \int_B \delta f_{\mathbf{x},\mathbf{y}}(\xi, \gamma - \mathbf{y}_0)\, d\gamma}{f_{\mathbf{y}}(\mathbf{y}_0) + (1/V_B) \int_B \delta f_{\mathbf{y}}(\rho - \mathbf{y}_0)\, d\rho} \tag{3-36}$$

We assume that $f_{\mathbf{x},\mathbf{y}}$ and $f_{\mathbf{y}}$ are continuous, so that the mean value theorem can be used to write

$$\int_B \delta f_{\mathbf{x},\mathbf{y}}\, d\gamma = V_B\, \delta f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{b}_1) \tag{3-37a}$$

$$\int_B \delta f_{\mathbf{y}}\, d\rho = V_B\, \delta f_{\mathbf{y}}(\mathbf{b}_2) \tag{3-37b}$$

for $\mathbf{b}_1$ and $\mathbf{b}_2$ vectors somewhere in the hypercube B. Thus,

$$f_{\mathbf{x}}(\xi \mid \mathbf{y}(\omega) \in B) = \frac{f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{y}_0) + \delta f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{b}_1)}{f_{\mathbf{y}}(\mathbf{y}_0) + \delta f_{\mathbf{y}}(\mathbf{b}_2)} \tag{3-38}$$

Now consider reducing the hypercube down to the point $\mathbf{y}_0$ by letting $l \to 0$. This causes $\mathbf{b}_1 \to \mathbf{y}_0$, $\mathbf{b}_2 \to \mathbf{y}_0$, $\delta f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{b}_1) \to 0$, and $\delta f_{\mathbf{y}}(\mathbf{b}_2) \to 0$, so that

$$\lim_{l \to 0} f_{\mathbf{x}}(\xi \mid \mathbf{y}(\omega) \in B) = f_{\mathbf{x},\mathbf{y}}(\xi, \mathbf{y}_0)/f_{\mathbf{y}}(\mathbf{y}_0) \tag{3-39}$$

defines what is meant by $f_{\mathbf{x}|\mathbf{y}}(\xi \mid \mathbf{y}_0)$, sometimes denoted as $f_{\mathbf{x}}(\xi \mid \mathbf{y} = \mathbf{y}_0)$. This development is not the most general possible, but is sufficient for the continuous random variable problems to be considered later.

A fundamental result, which some use as the basic definition of a *conditional probability density*, is:

$$f_{\mathbf{x}|\mathbf{y}}(\xi \mid \rho) = \frac{f_{\mathbf{x},\mathbf{y}}(\xi, \rho)}{f_{\mathbf{y}}(\rho)} \tag{3-40}$$

The denominator in (3-40) can be interpreted as a term to normalize the expression, so that the total "area under the density" is unity:

$$\int_{-\infty}^{\infty} f_{\mathbf{x}|\mathbf{y}}(\xi \mid \rho)\, d\xi = \int_{-\infty}^{\infty} \frac{f_{\mathbf{x},\mathbf{y}}(\xi, \rho)}{f_{\mathbf{y}}(\rho)}\, d\xi = \frac{\int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(\xi, \rho)\, d\xi}{f_{\mathbf{y}}(\rho)}$$

$$= \frac{f_{\mathbf{y}}(\rho)}{f_{\mathbf{y}}(\rho)} = 1 \tag{3-41}$$

A graphical representation of (3-40) is useful for insight. Figure 3.13 portrays the joint density function $f_{\mathrm{x},\mathrm{y}}(\xi, \rho)$ as a surface above the ξ–ρ plane. To generate $f_{\mathrm{x}|\mathrm{y}}(\xi \mid y_0)$ from this surface, a plane is passed through the surface at $\rho = y_0$, orthogonal to the ρ axis, resulting in the shaded region being $f_{\mathrm{x},\mathrm{y}}(\xi, y_0)$. If that

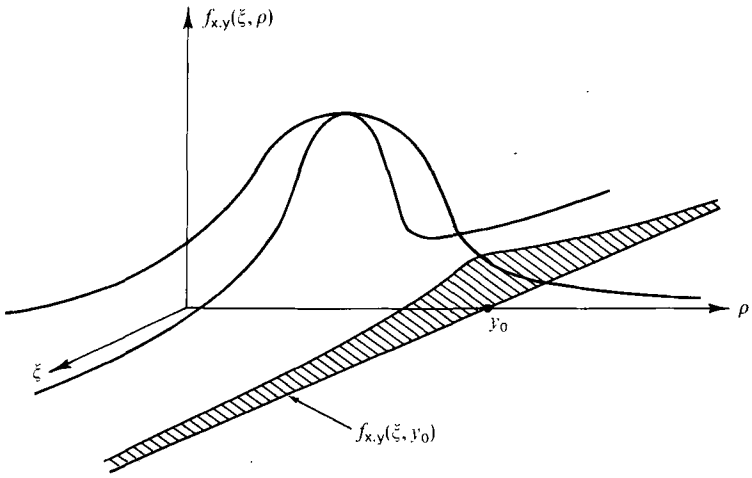

FIG. 3.13 Generation of conditional density.

function is divided by the *number* $f_y(y_0)$, the resulting function is normalized: its height is adjusted so that the shaded region is of area one. Note that $f_y(y_0)$ can be obtained from integrating $f_{x,y}(\xi, \rho)$ over all ξ, and evaluating the resulting function at $\rho = y_0$.

Eventually, we will want to consider the problem of estimating the value of a state vector $\mathbf{x}$ based upon a set of measurements $\mathbf{z}_1 = \mathbf{z}_1, \mathbf{z}_2 = \mathbf{z}_2, \ldots, \mathbf{z}_N = \mathbf{z}_N$. To accomplish this objective, we will propagate the conditional density $f_{\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N}(\xi | \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$ of the state $\mathbf{x}$ (modeled as a random variable), conditioned on knowledge of the entire set of measurements. This density embodies all of the information needed for estimation purposes.

Equation (3-40) is one form of *Bayes' rule*. Another useful form of this rule is

$$f_{\mathbf{x}|\mathbf{y}}(\xi | \rho) = \frac{f_{\mathbf{x},\mathbf{y}}(\xi, \rho)}{f_{\mathbf{y}}(\rho)} = \frac{f_{\mathbf{y}|\mathbf{x}}(\rho | \xi) f_{\mathbf{x}}(\xi)}{f_{\mathbf{y}}(\rho)} \tag{3-42}$$

Through this expression, one can readily generate the conditional density $f_{\mathbf{x}|\mathbf{y}}(\xi | \rho)$ if it is possible to write $f_{\mathbf{y}|\mathbf{x}}(\rho | \xi)$ and the unconditional densities for $\mathbf{x}$ and $\mathbf{y}$. This will be exploited to derive $f_{\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N}(\xi | \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$ for the estimation problem as just described. The denominator in Eq. (3-42) can be expanded to yield yet another useful form of Bayes' rule:

$$f_{\mathbf{x}|\mathbf{y}}(\xi | \rho) = \frac{f_{\mathbf{y}|\mathbf{x}}(\rho | \xi) f_{\mathbf{x}}(\xi)}{\int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(\zeta, \rho)\, d\zeta} = \frac{f_{\mathbf{y}|\mathbf{x}}(\rho | \xi) f_{\mathbf{x}}(\xi)}{\int_{-\infty}^{\infty} f_{\mathbf{y}|\mathbf{x}}(\rho | \zeta) f_{\mathbf{x}}(\zeta)\, d\zeta} \tag{3-43}$$

Note that the integrand in the denominator is of the same form as the numerator.

EXAMPLE 3.7 Consider two scalar random variables x and y, described through the joint density function

$$f_{x,y}(\xi, \rho) = (1/\pi)\exp\{-2[\xi - 1]^2 - 2[\rho - 2]^2 + 2\sqrt{3}[\xi - 1][\rho - 2]\}$$

The conditional probability density $f_{x|y}(\xi|\rho)$ can be generated from (3-40) once $f_y(\rho)$ is established through the concept of marginal densities:

$$f_y(\rho) = \int_{-\infty}^{\infty} f_{x,y}(\xi, \rho)\, d\xi = \frac{1}{\sqrt{2\pi}}\exp\{-\tfrac{1}{2}[\rho - 2]^2\}$$

Then, (3-40) yields, after some algebraic reduction,

$$f_{x|}\ (\xi|\rho) = (1/\pi)\exp\{-2[\xi - (\sqrt{3}/2)\rho + \sqrt{3} - 1]^2\}$$

This is the density function for x as a function of ξ, given that y has assumed some given realization ρ: given a particular ρ, the density is completely specified. ■

Through conditional probabilities and densities, we are specifying interrelationships among random variables. The two extremes of such relationships are independence and functional dependence, to be discussed next.

Consider two random variables, $\mathbf{x}$ mapping Ω into R^n and $\mathbf{y}$ mapping Ω into R^m, and two admissible events $A \subset R^n$ and $B \subset R^m$. Then $\mathbf{x}$ and $\mathbf{y}$ are *independent* if

$$P(\{\omega:\mathbf{x}(\omega) \in A \text{ and } \mathbf{y}(\omega) \in B\}) = P(\{\omega:\mathbf{x}(\omega) \in A\})P(\{\omega:\mathbf{y}(\omega) \in B\}) \quad (3\text{-}44)$$

for all A and B. This is a fundamental definition, in terms of sets in the sample space Ω. To relate it to distribution functions, let A and B be chosen as sets of a particular form:

$$A = \{\mathbf{x}:\mathbf{x} = \mathbf{x}(\omega) \leq \boldsymbol{\xi}\} = \{\mathbf{x}:x_1 \leq \xi_1, x_2 \leq \xi_2, \ldots, x_n \leq \xi_n\} \quad (3\text{-}45a)$$

$$B = \{\mathbf{y}:\mathbf{y} = \mathbf{y}(\omega) \leq \boldsymbol{\rho}\} \quad (3\text{-}45b)$$

Then, by the definition of the appropriate distribution functions,

$$P(\{\omega:\mathbf{x}(\omega) \in A \text{ and } \mathbf{y}(\omega) \in B\}) = F_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \boldsymbol{\rho}) \quad (3\text{-}46a)$$

$$P(\{\omega:\mathbf{x}(\omega) \in A\}) = F_{\mathbf{x}}(\boldsymbol{\xi}) \quad (3\text{-}46b)$$

$$P(\{\omega:\mathbf{y}(\omega) \in B\}) = F_{\mathbf{y}}(\boldsymbol{\rho}) \quad (3\text{-}46c)$$

for this particular choice of A and B. Thus, if $\mathbf{x}$ and $\mathbf{y}$ are independent, then

$$F_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \boldsymbol{\rho}) = F_{\mathbf{x}}(\boldsymbol{\xi})F_{\mathbf{y}}(\boldsymbol{\rho}) \quad (3\text{-}47)$$

for all $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$. If the distribution functions in (3-47) all have well-defined derivatives, one can conclude that, if $\mathbf{x}$ and $\mathbf{y}$ are independent, then

$$f_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \boldsymbol{\rho}) = f_{\mathbf{x}}(\boldsymbol{\xi})f_{\mathbf{y}}(\boldsymbol{\rho}) \quad (3\text{-}48)$$

for all $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$.

Another, more restrictive, approach is to define random vectors $\mathbf{x}$ and $\mathbf{y}$ to be independent if their joint density $f_{\mathbf{x},\mathbf{y}}(\xi, \rho)$ can be equated to the product of the separate marginal densities $f_{\mathbf{x}}(\xi)$ and $f_{\mathbf{y}}(\rho)$, as in (3-48). However, such a "definition" is valid only if the densities involved exist, whereas the concept of independence does not inherently require such existence.

When using the fundamental sample space and its subsets to think of independence, confusion sometimes arises between independent events and mutually exclusive events. If the occurrence of event A implies that B did not occur and vice versa, i.e., if $A \cap B = \varnothing$, then A and B are mutually exclusive. Said another way, if $P(A) = 1$ implies $P(B) = 0$ and $P(B) = 1$ implies $P(A) = 0$, then A and B are mutually exclusive. However, if knowledge of $P(A)$ gives you no information about $P(B)$ and vice versa, i.e., if $P(A \text{ and } B) = P(A)P(B)$, then A and B are independent events.

If $\mathbf{x}$ and $\mathbf{y}$ are independent, then (3-48) and Bayes' rule together yield $f_{\mathbf{x}|\mathbf{y}}(\xi \,|\, \rho)$ as

$$f_{\mathbf{x}|\mathbf{y}}(\xi \,|\, \rho) = f_{\mathbf{x},\mathbf{y}}(\xi, \rho)/f(\rho) = f_{\mathbf{x}}(\xi) f_{\mathbf{y}}(\rho)/f_{\mathbf{y}}(\rho) = f_{\mathbf{x}}(\xi) \tag{3-49}$$

i.e., the conditional density for $\mathbf{x}$, conditioned on knowledge that $\mathbf{y}$ has assumed a realization ρ, is equal to the unconditional density for $\mathbf{x}$. This makes sense conceptually: if $\mathbf{x}$ and $\mathbf{y}$ are to be independent, then knowledge of the value of $\mathbf{y}(\omega)$ should give no information about the value of $\mathbf{x}(\omega)$.

In practice, physical arguments are often presented to establish the independence of two random variables. The validity of such arguments can be established through empirical testing as well. For example, if x describes the outcome of one toss of a coin, and y models the outcome of another toss of a coin, intuition dictates that there is no causal relationship between x and y—that the outcome of one toss does not affect the other toss. Experimental testing can substantiate (or contradict) such intuition. In other cases, uncertainty in the value of two quantities can be ascribed to physically unrelated sources. For instance, consider a sampled signal from an aircraft inertial system, modeled as a true position indication corrupted by a noise random variable $\mathbf{n}_{\text{INS}}$, and a sample of ground-based radar data, modeled similarly as a true position indication corrupted by noise $\mathbf{n}_{\text{radar}}$. The sources of uncertainty modeled through $\mathbf{n}_{\text{INS}}$ (accelerometer bias, gyro drift, aircraft bending, etc.) are physically unrelated to the effects modeled by $\mathbf{n}_{\text{radar}}$ (electronic noise, atmospheric effects, etc.), and so these random variables are assumed to be independent.

The other extreme case of random variable interrelationship is *functional dependence*. If $\mathbf{x}$ is a deterministic function of $\mathbf{y}$, $\mathbf{x} = \boldsymbol{\phi}(\mathbf{y})$, then the conditional probability density function $f_{\mathbf{x}|\mathbf{y}}(\xi \,|\, \rho)$ is an impulse:

$$f_{\mathbf{x}|\mathbf{y}}(\xi \,|\, \rho) = \delta[\xi - \boldsymbol{\phi}(\rho)] = \delta[\boldsymbol{\phi}(\rho) - \xi] \tag{3-50}$$

where the delta function $\delta(\cdot)$ is defined as the function which satisfies the conditions:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(\boldsymbol{\xi})\, d\xi_1 \cdots d\xi_n = 1; \qquad \delta(\boldsymbol{\xi}) = 0 \qquad \text{for all} \quad \boldsymbol{\xi} \neq \mathbf{0} \tag{3-51}$$

It assumes a value of zero everywhere in R^n except where its argument is $\mathbf{0}$, and its integrated value over all R^n is unity. Equation (3-50) asserts that the conditional density function collapses down to an impulse function along the graph $\boldsymbol{\xi} = \boldsymbol{\phi}(\boldsymbol{\rho})$. If $\mathbf{x} = \boldsymbol{\phi}(\mathbf{y})$ and $\mathbf{y}(\omega) = \boldsymbol{\rho}$, then $\mathbf{x}(\omega) = \boldsymbol{\phi}[\mathbf{y}(\omega)] = \boldsymbol{\phi}(\boldsymbol{\rho})$ with no uncertainty: all of the probability is concentrated at $\boldsymbol{\xi} = \boldsymbol{\phi}(\boldsymbol{\rho})$, rather than being spread over a range of $\boldsymbol{\xi}$ values. In general, we will want to avoid impulse density functions, employing discontinuous distribution functions in such cases instead. However, the geometric insight of a density function collapsing down along certain loci of $\boldsymbol{\xi}$ values will be of practical use in certain applications, such as a system model outputting a number of "perfect" measurements.

3.5 FUNCTIONS OF RANDOM VARIABLES

The preceding section introduced the concept of functions of random variables, and this warrants some further attention. Let $\mathbf{x}$ be a vector random variable that maps the sample space Ω into n-dimensional Euclidean space R^n. Now consider a continuous mapping $\boldsymbol{\theta}(\cdot)$ from R^n into R^m, thus generating a vector $\mathbf{y} \in R^m$ from a vector $\mathbf{x} \in R^n$, as depicted in Fig. 3.14. Actually, $\boldsymbol{\theta}(\cdot)$ can be out of a larger class of functions than the continuous functions, called Baire functions (Borel measurable functions), composed of continuous functions and limits of continuous functions, but this generality will not be needed in our applications.

Now define the m-vector-valued function $\mathbf{y}$ as the composite mapping $\boldsymbol{\theta}[\mathbf{x}(\cdot)]$. Then $\mathbf{y}$ is itself a random variable, denoted as $\mathbf{y}$:

$$\mathbf{y}(\cdot) = \boldsymbol{\theta}[\mathbf{x}(\cdot)] \tag{3-52a}$$

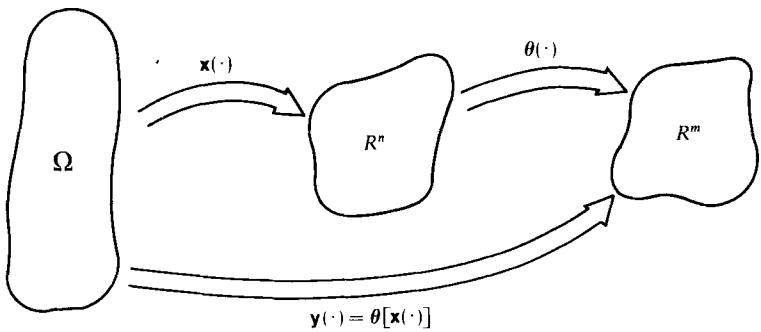

FIG. 3.14 Function of a random variable.

or

$$\begin{aligned} \mathsf{y}_1(\cdot) &= \theta_1[\mathsf{x}_1(\cdot), \mathsf{x}_2(\cdot), \ldots, \mathsf{x}_n(\cdot)] \\ &\vdots \\ \mathsf{y}_m(\cdot) &= \theta_m[\mathsf{x}_1(\cdot), \mathsf{x}_2(\cdot), \ldots, \mathsf{x}_n(\cdot)] \end{aligned} \tag{3-52b}$$

Stated simply, every Baire function of a random variable is a random variable.

Recall Eq. (3-28): $\mathbf{x}(\cdot)$ generates a new probability space $(R^n, \mathscr{F}_B, P_x)$ from the original probability space $(\Omega, \mathscr{F}, P)$, with P_x defined in (3-26) or (3-27). If $\boldsymbol{\theta}(\cdot)$ is a Baire function (Borel measurable) on R^n, then for every set of interest B in the range space R^m, the inverse image in R^n, $\{x \in R^n : \boldsymbol{\theta}(\mathbf{x}) \in B\}$, is an event for which probability has been defined through P_x. If we were to view $(R^n, \mathscr{F}_B, P_x)$ as the underlying probability space, then this just defines $\boldsymbol{\theta}(\cdot)$ itself as a random variable mapping from the sample space R^n into the space R^m.

Analogous to the discussion concerning (3-28), we would then expect to generate a new probability space, $(R^m, \mathscr{F}_B, P_y)$. The sets of interest in R^m will be the elements of the Borel field $\mathscr{F}_B$ associated with R^m, and for all sets $B \subset R^m$ and $B \in \mathscr{F}_B$, we can define probabilities through an appropriate probability function (measure), $P_y(\cdot)$, to be described shortly. Thus,

$$(\Omega, \mathscr{F}, P) \xrightarrow{\mathbf{x}(\cdot)} (R^n, \mathscr{F}_B, P_x) \xrightarrow{\boldsymbol{\theta}(\cdot)} (R^m, \mathscr{F}_B, P_y) \tag{3-53}$$

Then $\mathbf{y}(\cdot) = \boldsymbol{\theta}[\mathbf{x}(\cdot)]$ is a random variable that directly maps into this new probability space:

$$(\Omega, \mathscr{F}, P) \xrightarrow{\mathbf{y}(\cdot) = \boldsymbol{\theta}[\mathbf{x}(\cdot)]} (R^m, \mathscr{F}_B, P_y) \tag{3-54}$$

The random variable $\mathbf{y}$ has a *distribution induced by the distribution of* $\mathbf{x}$:

$$\begin{aligned} F_{\mathbf{y}}(\boldsymbol{\rho}) &= P(\{\omega : \mathbf{y}(\omega) \leq \boldsymbol{\rho}\}) = P(\{\omega : \boldsymbol{\theta}[\mathbf{x}(\omega)] \leq \boldsymbol{\rho}\}) \\ &= P_x(\{\mathbf{x} : \boldsymbol{\theta}(\mathbf{x}) \leq \boldsymbol{\rho}\}) \end{aligned} \tag{3-55}$$

The *induced probability density function*, if it exists, is given by

$$f_{\mathbf{y}}(\boldsymbol{\rho}) = \frac{\partial^m}{\partial \rho_1 \, \partial \rho_2 \cdots \partial \rho_m} F_{\mathbf{y}}(\boldsymbol{\rho}) \tag{3-56}$$

EXAMPLE 3.8 Consider the die-toss experiment discussed previously in Examples 3.3 and 3.5, in which we were interested only in the sets $A_1 = \{1 \text{ or } 2\}$ and $A_2 = \{3\}$. We defined a random variable $\mathsf{x}(\cdot)$ through

$$\mathsf{x}(\omega) = \begin{cases} 0 & \text{if } \omega \notin A_1 \text{ or } A_2 \\ 1 & \text{if } \omega \in A_1 \\ 2 & \text{if } \omega \in A_2 \end{cases}$$

and derived its probability distribution function.

Now let us say that you will receive a payoff according to what you roll on the die. Let the payoff function $\theta(\cdot)$ be defined as

$$\theta(x) = x^2 + 1$$

so that $\theta(\mathbf{x})$ is a random variable $\mathbf{y}$ defined as

$$\mathbf{y} = \theta(\mathbf{x}) = \mathbf{x}^2 + 1$$

Now we want to establish the induced distribution of $\mathbf{y}$, $F_{\mathbf{y}}(\rho)$, to describe our potential payoff. This can be generated using

$$F_{\mathbf{y}}(\rho) = P(\{\omega:\mathbf{y}(\omega) \leq \rho\}) = P(\{\omega:\theta[\mathbf{x}(\omega)] \leq \rho\}) = P(\{\omega:\mathbf{x}^2(\omega) + 1 \leq \rho\})$$

For $\rho < 1$, $P(\{\omega:\mathbf{x}^2(\omega) + 1 \leq \rho\}) = P(\varnothing) = 0$.
For $1 \leq \rho < 2$, $P(\{\omega:\mathbf{x}^2(\omega) + 1 \leq \rho\}) = P(\{A_1 \cup A_2\}^*) = \frac{1}{2}$.
For $2 \leq \rho < 5$, $P(\{\omega:\mathbf{x}^2(\omega) + 1 \leq \rho\}) = P(A_2^*) = \frac{5}{6}$.
For $5 \leq \rho < \infty$, $P(\{\omega:\mathbf{x}^2(\omega) + 1 \leq \rho\}) = P(\Omega) = 1$.

Thus we obtain the distribution function for $\mathbf{y}$ induced by the distribution of $\mathbf{x}$ as plotted in Fig. 3.15. ■

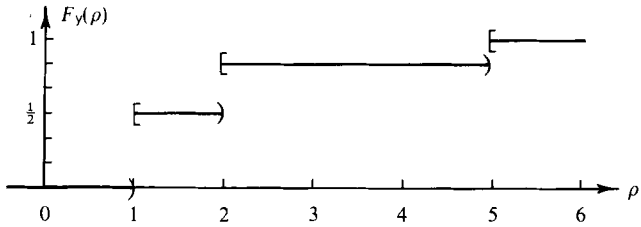

FIG. 3.15 Induced probability distribution function.

EXAMPLE 3.9 The scalar Gaussian random variable $\mathbf{x}$ is defined on $\Omega = R^1$ and is described through the density function

$$f_{\mathbf{x}}(\xi) = (1/\sqrt{2\pi P})\exp\{-(1/2P)(\xi - m)^2\}$$

Let the random variable $\mathbf{y}$ be defined through

$$\mathbf{y} = \theta(\mathbf{x}) = \mathbf{x}^3$$

Now we want to generate the density function to describe $\mathbf{y}$.

First, the distribution function is

$$\begin{aligned} F_{\mathbf{y}}(\rho) &= P(\{\omega:\mathbf{y}(\omega) \leq \rho\}) = P(\{\omega:\mathbf{x}^3(\omega) \leq \rho\}) \\ &= P_{\mathbf{x}}(\{x:x^3 \leq \rho\}) = P_{\mathbf{x}}(\{x:x \leq \rho^{1/3}\}) = F_{\mathbf{x}}(\rho^{1/3}) = \int_{-\infty}^{\rho} f_{\mathbf{x}}(\xi)\,d\xi \end{aligned}$$

Then the desired density function is the derivative of $F_{\mathbf{y}}(\rho)$ with respect to ρ. Using Leibnitz' rule, this yields

$$\begin{aligned} f_{\mathbf{y}}(\rho) &= \frac{dF_{\mathbf{y}}(\rho)}{d\rho} = \frac{d}{d\rho}\left\{\int_{-\infty}^{\rho^{1/3}} f_{\mathbf{x}}(\xi)\,d\xi\right\} \\ &= \frac{d}{d\rho}\{\rho^{1/3}\} \cdot f_{\mathbf{x}}(\rho^{1/3}) = \frac{1}{3}\rho^{-2/3} \cdot \frac{1}{\sqrt{2\pi P}}\exp\left\{-\frac{1}{2P}(\rho^{1/3} - m)^2\right\} \end{aligned}$$ ■

Thus, given a random variable $\mathbf{x}$ and its distribution (or density if it exists) and the functional relationship $\mathbf{y} = \boldsymbol{\theta}(\mathbf{x})$, the induced distribution (density) for $\mathbf{y}$ can be determined. If densities do exist, then a useful result can be sum-

marized in the following manner. Let $\mathbf{x}$ and $\mathbf{y}$ be n-dimensional vector random variables, with $\mathbf{y} = \boldsymbol{\theta}(\mathbf{x})$. Suppose $\boldsymbol{\theta}^{-1}$ exists and both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{-1}$ are continuously differentiable. Then

$$f_{\mathbf{y}}(\boldsymbol{\rho}) = f_{\mathbf{x}}[\boldsymbol{\theta}^{-1}(\boldsymbol{\rho})]\left\|\partial\boldsymbol{\theta}^{-1}(\boldsymbol{\rho})/\partial\boldsymbol{\rho}\right\| \tag{3-57}$$

where $\left\|\partial\boldsymbol{\theta}^{-1}(\boldsymbol{\rho})/\partial\boldsymbol{\rho}\right\| > 0$ is the absolute value of the Jacobian determinant arising naturally from integrations with change of variables. A proof of this theorem is outlined in Problem 3.8 at the end of this chapter.

EXAMPLE 3.10 Let us apply (3-57) to Example 3.9. Since $y = \theta(x) = x^3$, the inverse function θ^{-1} exists and can be expressed as (the real root)

$$\theta^{-1}(\rho) = \rho^{1/3}$$

and thus its derivative is

$$\partial\theta^{-1}(\rho)/\partial\rho = \tfrac{1}{3}\rho^{-2/3}$$

From (3-57),

$$\begin{aligned} f_y(\rho) &= f_x(\rho^{1/3}) \cdot \tfrac{1}{3}\rho^{-2/3} \\ &= \tfrac{1}{3}\rho^{-2/3}(1/\sqrt{2\pi P})\exp\{-(1/2P)(\rho^{1/3} - m)^2\} \end{aligned}$$

as found previously. ■

Now consider the set function P_y introduced earlier. As in (3-26), for all sets $B \subset R^m$ and $B \in \mathscr{F}_{\mathbf{B}}$, we can define probabilities as

$$P_y(B) = \int_B dF_{\mathbf{y}}(\boldsymbol{\rho}) \tag{3-58}$$

where meaning is given to the right hand side through measure theory. If density functions exist,

$$P_y'(B) = \int_B f_{\mathbf{y}}(\boldsymbol{\rho})\, d\boldsymbol{\rho} = P_y(B) \tag{3-59}$$

The idea of induced densities and distributions is then embodied in the following result. If $\mathbf{x}$ is a random variable mapping Ω into R^n, and $\mathbf{y} = \boldsymbol{\theta}(\mathbf{x})$, where $\boldsymbol{\theta}$ maps R^n into R^m, then

$$P_y(B) = P_x(\{\mathbf{x}:\boldsymbol{\theta}(\mathbf{x}) \in B\}) = P(\{\omega:\boldsymbol{\theta}[\mathbf{x}(\omega)] \in B\}) \tag{3-60}$$

The discussion in this section can help prevent the confusion that often arises in estimation. Consider having measurements available with which you want to estimate some quantities of interest, denoted as the n-dimensional vector $\boldsymbol{\theta}$. Suppose that the measurements are modeled through an m-dimensional vector of random variables $\mathbf{z}(\cdot)$, so that the numbers coming from the measuring devices are the realizations of the random variables for a particular outcome $\omega:\mathbf{z}(\omega) = \mathbf{z}$. Now you want to generate a mapping $\hat{\boldsymbol{\theta}}(\cdot)$ from R^m into R^n, called an estimator, that will map the given realizations into a "best" estimate of the value of $\boldsymbol{\theta}$. The composite mapping $\hat{\boldsymbol{\theta}}[\mathbf{z}(\cdot)]$ can then be considered a random

variable, often denoted as $\hat{\boldsymbol{\theta}}$, called a randomized estimate or estimator, or just an estimate. Finally, estimation algorithms generate vectors of numbers, $\hat{\boldsymbol{\theta}}(\mathbf{z}) \in R^n$, which are also called estimates. Whether one is concerned with a functional mapping, a random variable, or a vector of numbers is a fundamental distinction to be made, but one that can be misinterpreted unless one takes care to be aware of this aspect. The notation adopted herein specifically attempts to clarify this issue.

3.6 EXPECTATION AND MOMENTS OF RANDOM VARIABLES

The distribution or density function for a random variable is the entity of fundamental interest in Bayesian estimation, embodying all information known about the variable. Once it is generated, an "optimal" estimate can be defined using some chosen criterion. Similarly, it can be used to compute the expected value of some function of the random variable, where this "expected value" is just the average value one would obtain over the ensemble of outcomes of an "experiment." The expected value of particular functions will generate moments of a random variable, which are parameters (statistics) that characterize the distribution or density function. Although one would like to portray these functions completely through estimation, it is generally more feasible to evaluate expressions for a finite number of moments instead, thereby generating a partial description of the functions. In the case of Gaussian random variables, it will turn out that specification of only the first two moments will *completely* describe the distribution or density function.

Let $\mathbf{x}$ be an n-dimensional random variable vector described through a density function $f_{\mathbf{x}}(\xi)$, and let $\mathbf{y}$ be an m-dimensional vector function of $\mathbf{x}$:

$$\mathbf{y}(\cdot) = \boldsymbol{\theta}[\mathbf{x}(\cdot)] \tag{3-61}$$

where $\boldsymbol{\theta}(\cdot)$ is continuous. Thus $\mathbf{y}$ is also a random variable with induced density $f_{\mathbf{y}}(\boldsymbol{\rho})$. Then the *expectation* of $\mathbf{y}$ is

$$E[\mathbf{y}] = \int_{-\infty}^{\infty} \boldsymbol{\theta}(\boldsymbol{\xi}) f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi} = \int_{-\infty}^{\infty} \boldsymbol{\rho} f_{\mathbf{y}}(\boldsymbol{\rho})\, d\boldsymbol{\rho} \tag{3-62}$$

If the density function $f_{\mathbf{x}}(\xi)$ does not exist, then $E[\mathbf{y}]$ can still be defined as

$$E[\mathbf{y}] = \int_{\Omega} \boldsymbol{\theta}[\mathbf{x}(\omega)]\, dP(\omega) = \int_{R^n} \boldsymbol{\theta}(\boldsymbol{\xi})\, dF_{\mathbf{x}}(\boldsymbol{\xi}) = \int_{R^m} \boldsymbol{\rho}\, dF_{\mathbf{y}}(\boldsymbol{\rho}) \tag{3-63}$$

using measure theory to give meaning to the indicated integrals [12, 13, 15]. Note that the succeeding integrals in (3-63) are carried out over Ω, R^n, and R^m, respectively, using the appropriate probability functions. The integration performed over the original sample space Ω naturally provides the most basic

definition of expectation. However, for our applications, we will assume $f_{\mathbf{x}}(\xi)$ exists, so (3-62) will suffice as a definition.

EXAMPLE 3.11 Let us calculate the expected payoff for the die-toss experiment described in Example 3.8. A distribution function as in Fig. 3.15 indicates that the random variable assumes only discrete point values, with probability equal to the magnitude of each associated discontinuity [see Eq. (3-14)]. In such a case, the integrals indicated in (3-63) become summations, as

$$E[\mathsf{y}] = \int \rho \, dF_{\mathsf{y}}(\rho) = \sum_i \rho_i \Delta F_{\mathsf{y}}(\rho_i) = 1 \cdot (1 - P_1 - P_2) + 2 \cdot P_1 + 5 \cdot P_2$$

$$= 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{3} + 5 \cdot \tfrac{1}{6} = 2$$

A distribution function that varies continuously except for a finite number of jump discontinuities can be decomposed into the sum of a continuous function and a function composed only of the jump discontinuities as in Fig. 3.15. This allows expectations to be evaluated through addition of results obtained from the separate functions. ■

Since expectation is by definition an integration, it is a linear operation. In other words, for c equal to a scalar constant,

$$E[c\mathbf{y}] = cE[\mathbf{y}] \tag{3-64a}$$

$$E[\mathbf{y}_1 + \mathbf{y}_2] = E[\mathbf{y}_1] + E[\mathbf{y}_2] \tag{3-64b}$$

Combining (3-64a) and (3-64b) yields the useful result that, for $\mathbf{A}$ a known matrix,

$$E[\mathbf{Ay}] = \mathbf{A}E[\mathbf{y}] \tag{3-65}$$

Now let us consider some specific functions $\boldsymbol{\theta}(\cdot)$. First let $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{x}$ to generate the *first moment* of $\mathbf{x}$ or the *mean* of $\mathbf{x}$. Define an n-dimensional vector $\mathbf{m}$, whose components are the mean values $m_i \triangleq E[\mathsf{x}_i]$:

$$\mathbf{m} \triangleq \begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix} \triangleq \begin{bmatrix} E[\mathsf{x}_1] \\ \vdots \\ E[\mathsf{x}_n] \end{bmatrix} = \begin{bmatrix} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_1 f_{\mathbf{x}}(\xi) \, d\xi_1 \cdots d\xi_n \\ \vdots \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_n f_{\mathbf{x}}(\xi) \, d\xi_1 \cdots d\xi_n \end{bmatrix} \tag{3-66}$$

This vector is the mean or first moment of $\mathbf{x}$, and notationally (3-66) can be written equivalently as

$$\mathbf{m} \triangleq E[\mathbf{x}] = \int_{-\infty}^{\infty} \xi f_{\mathbf{x}}(\xi) \, d\xi \tag{3-67}$$

Next consider Eq. (3-62), letting $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{x}\mathbf{x}^{\mathrm{T}}$:

$$\mathbf{y} = \boldsymbol{\theta}(\mathbf{x}) = \mathbf{x}\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} \mathsf{x}_1^2 & \mathsf{x}_1\mathsf{x}_2 & \cdots & \mathsf{x}_1\mathsf{x}_n \\ \vdots & \vdots & & \vdots \\ \mathsf{x}_n\mathsf{x}_1 & \mathsf{x}_n\mathsf{x}_2 & \cdots & \mathsf{x}_n^2 \end{bmatrix} \tag{3-68}$$

Define a matrix, denoted as Ψ, as the n-by-n matrix whose i–j component is the *correlation* of x_i and x_j (and thus the diagonal terms are autocorrelations, or mean squared values, the square roots of which are termed root mean squared, or *RMS*, values):

$$\Psi_{ij} \triangleq E[\mathsf{x}_i \mathsf{x}_j] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_i \xi_j f_{\mathbf{x}}(\boldsymbol{\xi})\, d\xi_1 \cdots d\xi_n \tag{3-69}$$

This matrix is the *second (noncentral) moment* of $\mathbf{x}$ or the *autocorrelation matrix of* $\mathbf{x}$, and can be written as

$$\boldsymbol{\Psi} \triangleq E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \int_{-\infty}^{\infty} \boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi} \tag{3-70}$$

where again this simply is a compact notation to be interpreted in the light of Eq. (3-69).

Let us consider yet another function, $\boldsymbol{\theta}(\mathbf{x}) = [(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}]$. This allows us to define an n-by-n matrix $\mathbf{P}$ whose i–j component is the *covariance* of x_i and x_j:

$$\begin{aligned} P_{ij} &\triangleq E[(\mathsf{x}_i - m_i)(\mathsf{x}_j - m_j)] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\xi_i - m_i)(\xi_j - m_j) f_{\mathbf{x}}(\boldsymbol{\xi})\, d\xi_1 \cdots d\xi_n \end{aligned} \tag{3-71}$$

Note specifically that the mean values $m_i = E[\mathsf{x}_i]$ and $m_j = E[\mathsf{x}_j]$ in (3-70) are *not* random variables, but are statistics: deterministic numbers. The matrix $\mathbf{P}$ is the *second central moment* of $\mathbf{x}$ or the *covariance matrix* of $\mathbf{x}$, and can be written as

$$\mathbf{P} \triangleq E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}] = \int_{-\infty}^{\infty} (\boldsymbol{\xi} - \mathbf{m})(\boldsymbol{\xi} - \mathbf{m})^{\mathrm{T}} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi} \tag{3-72}$$

In cases where there might be ambiguity as to what correlation or covariance matrix is being discussed, subscripts will be employed in the notation, such as $\mathbf{P}_{xx}$ or $\boldsymbol{\Psi}_{yy}$.

Because the covariance will be significant in our work, it will be characterized further. The matrix $\mathbf{P}$ is a symmetric, positive semidefinite matrix (its eigenvalues are nonnegative). The variances of the separate components of $\mathbf{x}$ are along the diagonal:

$$P_{ii} \triangleq E[(\mathsf{x}_i - m_i)^2] \tag{3-73}$$

The square root of a variance P_{ii} is termed the *standard deviation* of x_i, denoted as σ_i. Thus, the diagonal terms can be expressed as

$$P_{ii} \triangleq \sigma_i^2 \tag{3-74}$$

The *correlation coefficient* of x_i and x_j, denoted as r_{ij}, is defined as the ratio

$$r_{ij} \triangleq \frac{E[(\mathsf{x}_i - m_i)(\mathsf{x}_j - m_j)]}{(E[(\mathsf{x}_i - m_i)^2])^{1/2}(E[(\mathsf{x}_j - m_j)^2])^{1/2}} \triangleq \frac{P_{ij}}{\sigma_i \sigma_j} \tag{3-75}$$

Using (3-74) and (3-75), the covariance matrix $\mathbf{P}$ can be written as

$$\mathbf{P} = \begin{bmatrix} \sigma_1^2 & r_{12}\sigma_1\sigma_2 & \cdots & r_{1n}\sigma_1\sigma_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & r_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{bmatrix} \tag{3-76}$$

If the correlation coefficient r_{ij} is zero, then the components x_i and x_j are said to be *uncorrelated.* Consequently, if $\mathbf{P}$ is diagonal, i.e., if $r_{ij} = 0$ for all i and j with $i \neq j$, then $\mathbf{x}$ is said to be composed of uncorrelated components.

Another expression for the covariance matrix can be derived in the following manner, using the facts that $E[\cdot]$ is linear and the mean vector $\mathbf{m}$ is not random:

$$\begin{aligned} \mathbf{P} &= E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}] = E[\mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathbf{x}\mathbf{m}^{\mathrm{T}} - \mathbf{m}\mathbf{x}^{\mathrm{T}} + \mathbf{m}\mathbf{m}^{\mathrm{T}}] \\ &= E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - E[\mathbf{x}\mathbf{m}^{\mathrm{T}}] - E[\mathbf{m}\mathbf{x}^{\mathrm{T}}] + E[\mathbf{m}\mathbf{m}^{\mathrm{T}}] \\ &= E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - E[\mathbf{x}]\mathbf{m}^{\mathrm{T}} - \mathbf{m}E[\mathbf{x}^{\mathrm{T}}] + \mathbf{m}\mathbf{m}^{\mathrm{T}} \\ &= E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbf{m}\mathbf{m}^{\mathrm{T}} - \mathbf{m}\mathbf{m}^{\mathrm{T}} + \mathbf{m}\mathbf{m}^{\mathrm{T}} \\ \mathbf{P} &= E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbf{m}\mathbf{m}^{\mathrm{T}} \end{aligned} \tag{3-77}$$

This equation then directly relates the central and noncentral second moments. In the scalar case, it reduces to

$$P = E[\mathsf{x}^2] - (E[\mathsf{x}])^2 \tag{3-78}$$

The special cases of expectation that have been considered are just the first two moments of a random variable. (See Problems 3.24–3.27 for means of establishing best estimates of these moments from a finite set of empirical data.) Of course, there are higher ordered moments that can be used to characterize a probability density (or distribution) function. The mean relates where the density is centered, and the covariance gives an indication of the spread of the density about that mean value. In general, an endless number of moments would be required to specify a density function completely. In the particular case of a Gaussian random variable, the mean and covariance *completely* specify the density. By knowing the mean and covariance of a Gaussian random variable, you know *all* of the probability information contained in the associated density function, not just two parameters that partially describe its shape. This will be exploited to a great extent in linear estimation problems.

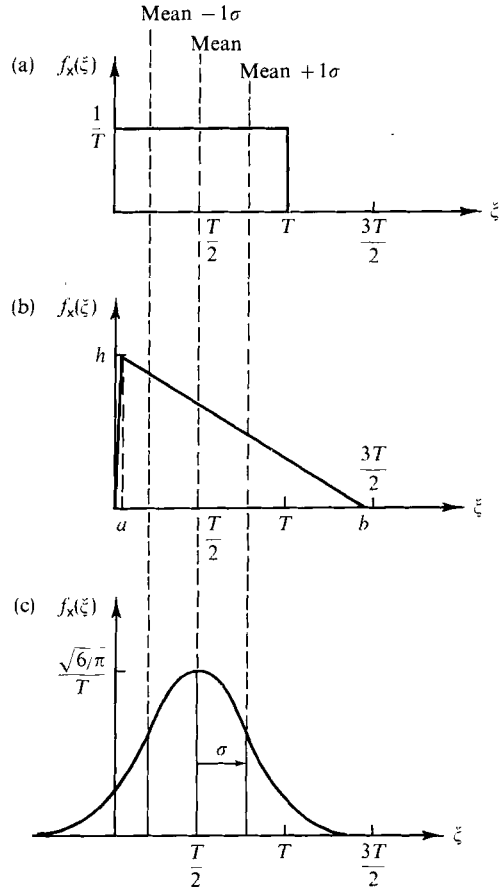

FIG. 3.16 Different random variables with equivalent first two moments. (a) Uniform. (b) Triangular. $a = (T/4)(3 - 5\sqrt{1/3}) \cong 0.03\ T$, $b = (T/4)(3 + 5\sqrt{1/3}) \cong 1.47\ T$, $h = (2/b) \cong 1.36\ T$. (c) Gaussian. $\sigma = T/\sqrt{12} \cong 0.29\ T$.

EXAMPLE 3.12 Consider the random variable x with uniform density between 0 and T, $f_x(\xi) = \{1/T \text{ for } \xi \in [0,T], 0 \text{ elsewhere}\}$, as depicted in Fig. 3.16a. The mean of x is

$$E[\mathsf{x}] = \int_{-\infty}^{\infty} \xi f_x(\xi)\, d\xi = \int_0^T \xi \frac{1}{T} d\xi = \frac{T}{2}$$

and the variance of x is

$$E\left[\left(\mathsf{x} - \frac{T}{2}\right)^2\right] = \int_0^T \left(\xi - \frac{T}{2}\right)^2 \frac{1}{2} d\xi = \frac{T^2}{12}$$

If y is defined as sin x, then $E[\mathsf{y}]$ is

$$E[\mathsf{y}] = E[\sin \mathsf{x}] = \int_0^T \sin \xi \frac{1}{T} d\xi = \frac{1}{T}(1 - \cos T)$$

Note that specification of just the first two moments of a random variable does not completely describe the associated distribution or density function. The triangular-shaped density function in Fig. 3.16b yields the same first two moments, despite its significantly different shape. Figure 3.16c depicts a Gaussian density yielding the same first two moments as in (a) and (b),

$$f_x(\xi) = \frac{1}{[2\pi(T^2/12)]^{1/2}} \exp\left\{-\frac{1}{2(T^2/12)}\left(\xi - \frac{T}{2}\right)^2\right\}$$

Furthermore, note that if one knew $f_x(\xi)$ were uniform or Gaussian, then knowledge of the first two moments would specify $f_x(\xi)$ completely. However, if $f_x(\xi)$ were triangular, these two parameters would not specify the density shape totally. An infinite number of moments is required to specify the shape of a general density function. ■

It will be useful to generalize the concept of the second moment of a single random variable $\mathbf{x}$ to the second moment relationship between two random variables $\mathbf{x}$ and $\mathbf{y}$. Such a concept is inherently involved in Eqs. (3-70) and (3-72), since the i–j component of Ψ or $\mathbf{P}$ is the cross-correlation or covariance, respectively, of the scalar random variables x_i and x_j; now we will generalize from the scalar to vector case. Let $\mathbf{x}$ be an n-dimensional random vector and $\mathbf{y}$ an m-dimensional random vector. Then the *cross-correlation matrix* of $\mathbf{x}$ and $\mathbf{y}$ is the n-by-m matrix whose i–j component is

$$E[\mathsf{x}_i\mathsf{y}_j] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_i \rho_j f_{\mathbf{x},\mathbf{y}}(\xi, \rho) d\xi_1 \cdots d\xi_n d\rho_1 \cdots d\rho_m \tag{3-79}$$

This matrix is then expressed notationally as

$$\Psi_{xy} \triangleq E[\mathbf{x}\mathbf{y}^{\mathrm{T}}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi \rho^{\mathrm{T}} f_{\mathbf{x},\mathbf{y}}(\xi, \rho)\, d\xi\, d\rho \tag{3-80}$$

Similarly, the second central moment generalizes to the *cross-covariance matrix* $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{P}_{xy} \triangleq E[(\mathbf{x} - \mathbf{m}_x)(\mathbf{y} - \mathbf{m}_y)^{\mathrm{T}}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\xi - \mathbf{m}_x)(\rho - \mathbf{m}_y)^{\mathrm{T}} f_{\mathbf{x},\mathbf{y}}(\xi, \rho)\, d\xi\, d\rho \tag{3-81}$$

Two random vectors $\mathbf{x}$ and $\mathbf{y}$ are termed *uncorrelated* if their correlation matrix is equal to the outer product of their first order moments, i.e., if

$$E[\mathbf{x}\mathbf{y}^{\mathrm{T}}] = E[\mathbf{x}]E[\mathbf{y}^{\mathrm{T}}] = \mathbf{m}_x\mathbf{m}_y^{\mathrm{T}} \tag{3-82a}$$

or

$$E[\mathsf{x}_i\mathsf{y}_j] = E[\mathsf{x}_i]E[\mathsf{y}_j] \qquad \text{for all} \quad i \text{ and } j \tag{3-82b}$$

which is equivalent to the condition that $E\{[\mathsf{x}_i - m_{x_i}][\mathsf{y}_j - m_{y_j}]\} = 0$ for all i and j.

EXAMPLE 3.13 We want to show by a simple example that the preceding definition of uncorrelatedness corresponds to the previous definition of uncorrelated scalar random variables which involved the correlation coefficient described by Eq. (3-75). Consider two scalar random

variables, z_1 and z_2. By (3-82), they are uncorrelated if $E[z_1z_2] = E[z_1]E[z_2]$. Now let $\mathbf{z}$ be the vector random variable made up of components z_1 and z_2. The covariance of $\mathbf{z}$ is then

$$\mathbf{P}_{zz} = \begin{bmatrix} E[z_1^2] - E[z_1]^2 & E[z_1z_2] - E[z_1]E[z_2] \\ E[z_1z_2] - E[z_1]E[z_2] & E[z_2^2] - E[z_2]^2 \end{bmatrix}$$

But, if z_1 and z_2 are uncorrelated, then $E[z_1z_2] - E[z_1]E[z_2] = 0$, and the off-diagonal terms are zero. This is just the condition of the correlation coefficient of z_1 and z_2 being zero, as described earlier. ■

Whereas uncorrelatedness is a condition under which generalized second moments can be expressed as products of first order moments, independence is a condition under which the entire joint distribution or density function can be expressed as a product of marginal functions. As might be expected then, if $\mathbf{x}$ and $\mathbf{y}$ are independent, then they are uncorrelated, but not necessarily vice versa. This implication can be expressed simply as

$$\mathbf{x} \text{ and } \mathbf{y} \text{ independent} \rightarrow \mathbf{x} \text{ and } \mathbf{y} \text{ uncorrelated} \tag{3-83}$$

This can be demonstrated readily: by definition, we can write

$$E[\mathbf{x}\mathbf{y}^T] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \boldsymbol{\xi}\boldsymbol{\rho}^T f_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \boldsymbol{\rho}) d\boldsymbol{\xi}\, d\boldsymbol{\rho}$$

If $\mathbf{x}$ and $\mathbf{y}$ are independent, then this becomes:

$$E[\mathbf{x}\mathbf{y}^T] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \boldsymbol{\xi}\boldsymbol{\rho}^T f_{\mathbf{x}}(\boldsymbol{\xi}) f_{\mathbf{y}}(\boldsymbol{\rho}) d\boldsymbol{\xi}\, d\boldsymbol{\rho}$$

Separating the integration yields the desired result:

$$E[\mathbf{x}\mathbf{y}^T] = \int_{-\infty}^{\infty} \boldsymbol{\xi} f_{\mathbf{x}}(\boldsymbol{\xi}) d\boldsymbol{\xi} \int_{-\infty}^{\infty} \boldsymbol{\rho}^T f_{\mathbf{y}}(\boldsymbol{\rho}) d\boldsymbol{\rho} = E[\mathbf{x}]E[\mathbf{y}^T]$$

If $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated, they are *not* necessarily independent. A counter-example to such an implication is given in the following example.

EXAMPLE 3.14 (modified from [3]) Let z be uniformly distributed between 0 and 1:

$$f_z(\zeta) = \begin{cases} 1 & \zeta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Now define x and y as $x = \sin(2\pi z)$ and $y = \cos(2\pi z)$. It will now be shown that x and y are uncorrelated, but not independent. They are uncorrelated since

$$E[x] = E[y] = E[xy] = E[x]E[y] = 0$$

However, consider higher order moments, as the fourth generalized moment:

$$E[x^2y^2] = \tfrac{1}{8}$$

But,

$$E[x^2] = E[y^2] = \tfrac{1}{2}$$

so that

$$E[x^2]E[y^2] = \tfrac{1}{4}$$

which is not equal to $E[x^2y^2]$. If x and y were independent, these would be equal. ■

Another related concept is that of orthogonality. Two random vectors $\mathbf{x}$ and $\mathbf{y}$ are termed *orthogonal* if their correlation matrix is the zero matrix: if $E[\mathbf{x}\mathbf{y}^T] = \mathbf{0}$. Obviously, this concept is interrelated with $\mathbf{x}$ and $\mathbf{y}$ being uncorrelated, and this relation is as follows. If either $\mathbf{x}$ or $\mathbf{y}$ (or both) is zero-mean, then orthogonality and uncorrelatedness of $\mathbf{x}$ and $\mathbf{y}$ imply each other. However, if neither is zero-mean, then $\mathbf{x}$ and $\mathbf{y}$ may be uncorrelated or orthogonal or neither, but they cannot be both orthogonal and uncorrelated. Orthogonality provides one means of defining an optimal estimate: if we generate an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ based on measurement data $\mathbf{z}$, then that estimate can be termed optimal if the error $(\mathbf{x} - \hat{\mathbf{x}})$ is orthogonal to the data. This geometrical concept is instrumental in deriving optimal estimators by means of "orthogonal projections," the original means of derivation of the Kalman filter. We, however, will employ a Bayesian approach to estimation in the sequel.

3.7 CONDITIONAL EXPECTATIONS

The concept of expectation of some function of random variables answers the question, if we were to conduct a large (endless) number of experiments, what average value (over the entire ensemble of experimental outcomes, $\omega \in \Omega$) of that function would we achieve? Conditional expectations provide the same information, but incorporate insights into occurrence of events in Ω gained through observations of realizations of related random variables. In this section, we will first define conditional expectation under the assumption that the appropriate conditional density function exists. After investigating its properties and applications, the definition will be generalized to allow consideration of this concept without such an assumption.

Let $\mathbf{x}$ and $\mathbf{y}$ be random variables mapping Ω into R^n and R^m, respectively, and let $\mathbf{z}$ be a continuous (Baire) function of $\mathbf{x}$,

$$\mathbf{z}(\cdot) = \boldsymbol{\theta}[\mathbf{x}(\cdot)] \tag{3-84}$$

so that $\mathbf{z}$ is itself a random variable mapping Ω into R^r. Then the *conditional expected value*, or *conditional mean*, of $\mathbf{z}$, conditioned on the fact that $\mathbf{y}$ has assumed the realization $\mathbf{y} \in R^m$, i.e., $\mathbf{y}(\omega) = \mathbf{y}$, is

$$E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \boldsymbol{\theta}(\boldsymbol{\xi}) f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\xi}|\mathbf{y})\, d\xi_1 \cdots d\xi_n \tag{3-85a}$$

$$= \int_{-\infty}^{\infty} \boldsymbol{\theta}(\boldsymbol{\xi}) f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\xi}|\mathbf{y})\, d\boldsymbol{\xi} \tag{3-85b}$$

The subscript $\mathbf{x}$ on $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}]$ denotes that the expectation operation (integration) is performed over the possible values of $\mathbf{x}$, and sometimes this subscript is not included in the notation. For a given value $\mathbf{y} \in R^m$, $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}]$ is a vector in R^r. Thus, $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \cdot]$ is a mapping from R^m into R^r, a function of the values $\mathbf{y} \in R^m$. Recall Section 3.5, Functions of Random Variables. If these $\mathbf{y}$ values are realizations of the random variable $\mathbf{y}$, then the conditional

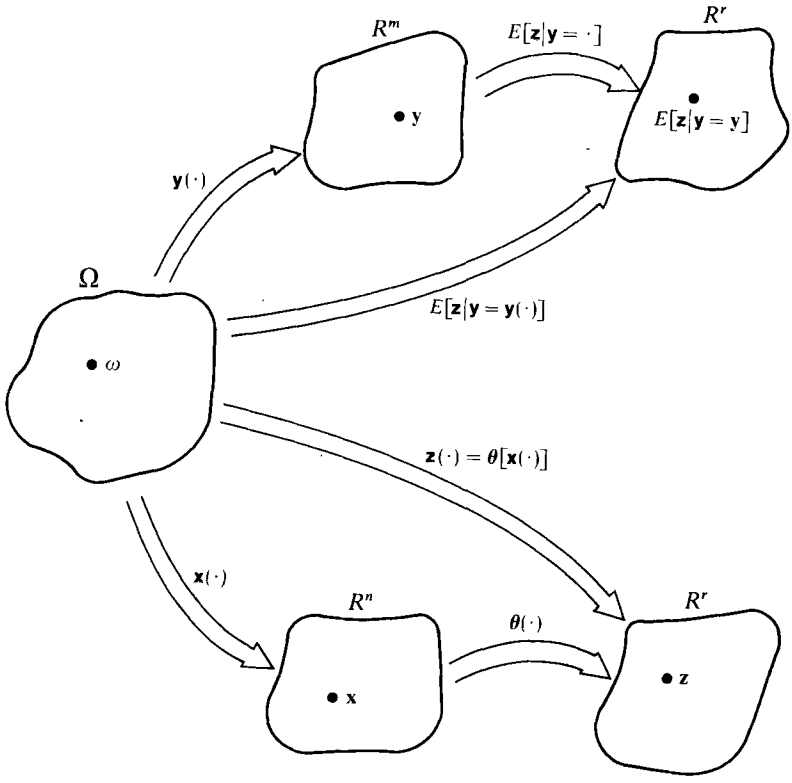

FIG. 3.17 Conditional expectation functional relationships.

expectation can be viewed as a random variable, i.e., the composite mapping $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}(\cdot)]$ mapping Ω into R^r. These interrelationships are depicted in Fig. 3.17.

Moreover, the random variable $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}(\cdot)]$ is unique and has the property that

$$E_{\mathbf{y}}\{E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}(\cdot)]\} = E_{\mathbf{x}}[\mathbf{z}] \tag{3-86}$$

Conceptually, this is reasonable. If we take the conditional expectation of $\mathbf{z}$, conditioned on a realized value of $\mathbf{y}$, and look at its expected value over all possible realizations of $\mathbf{y}$, then the result is the unconditional expectation of $\mathbf{z}$. Let us demonstrate the validity of (3-86) mathematically as well. By the definition of expectation, we can write

$$E_{\mathbf{x}}[\mathbf{z}] = \int_{-\infty}^{\infty} \theta(\xi) f_{\mathbf{x}}(\xi)\, d\xi$$

Now $f_{\mathbf{x}}(\xi)$ can be written as the marginal density derived from $f_{\mathbf{x},\mathbf{y}}(\xi, \rho)$ to yield

$$E_{\mathbf{x}}[\mathbf{z}] = \int_{-\infty}^{\infty} \theta(\xi) \left[\int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(\xi, \rho)\, d\rho \right] d\xi$$

Bayes' rule can be applied to derive an equivalent expression as

$$E_{\mathbf{x}}[\mathbf{z}] = \int_{-\infty}^{\infty} \theta(\xi)\left[\int_{-\infty}^{\infty} f_{\mathbf{x}|\mathbf{y}}(\xi|\rho) f_{\mathbf{y}}(\rho)\, d\rho\right] d\xi$$

Now we assume convergence in the definition of the integrals taken in different orders, so that we can interchange the order of integration (to be more precise, we invoke the Fubini theorem [7, 12, 13] from functional analysis) to yield

$$E_{\mathbf{x}}[\mathbf{z}] = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \theta(\xi) f_{\mathbf{x}|\mathbf{y}}(\xi|\rho)\, d\xi\right] f_{\mathbf{y}}(\rho)\, d\rho$$

The bracketed term is a function of ρ alone, where ρ is a dummy variable corresponding to realized values of $\mathbf{y}$ (ρ is used as the dummy variable to distinguish it from a single realization $\mathbf{y}$ of $\mathbf{y}$). Thus, the preceding expression is just the expected value (over all possible $\mathbf{y}$ realizations, ρ) of the bracketed term, which is itself the conditioned expectation of $\mathbf{z}$: this directly yields

$$E_{\mathbf{x}}[\mathbf{z}] = E_{\mathbf{y}}\{E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}(\cdot)]\}$$

as desired.

The conditional expectation can also be viewed as a function $E_{\mathbf{x}}[\cdot|\mathbf{y} = \mathbf{y}]$ that maps a random variable $\mathbf{z}$ into a vector $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}] \in R^r$. As in the case of unconditional expectations, such an operation is defined through an integration and is *linear*. Thus, if $\mathbf{A}$ is a known matrix,

$$E_{\mathbf{x}}[\mathbf{A}\mathbf{x}|\mathbf{y} = \mathbf{y}] = \mathbf{A}E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}] \tag{3-87}$$

$$E_{\mathbf{x}\mathbf{y}}[\mathbf{x} + \mathbf{y}|\mathbf{z} = \mathbf{z}] = E_{\mathbf{x}}[\mathbf{x}|\mathbf{z} = \mathbf{z}] + E_{\mathbf{y}}[\mathbf{y}|\mathbf{z} = \mathbf{z}] \tag{3-88}$$

Two special cases of the conditional mean of $\mathbf{z}$ defined in (3-85) are of particular interest to our applications: the conditional mean and covariance of $\mathbf{x}$. The *conditional mean* of $\mathbf{x}$, given that $\mathbf{y}$ has assumed the value $\mathbf{y}$, is generated by letting $\theta(\mathbf{x}) = \mathbf{x}$:

$$E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}] = \int_{-\infty}^{\infty} \xi f_{\mathbf{x}|\mathbf{y}}(\xi|\mathbf{y})\, d\xi \tag{3-89}$$

The *conditional covariance* of $\mathbf{x}$, given that $\mathbf{y}(\omega) = \mathbf{y}$, is then defined as

$$\mathbf{P}_{x|y} = E_{\mathbf{x}}\left[(\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}])(\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}])^{\mathrm{T}}\middle|\mathbf{y} = \mathbf{y}\right] \tag{3-90a}$$

$$= \int_{-\infty}^{\infty} (\xi - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}])(\xi - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}])^{\mathrm{T}} f_{\mathbf{x}|\mathbf{y}}(\xi|\mathbf{y})\, d\xi \tag{3-90b}$$

If we want to generate an estimate of $\mathbf{x}$ using measurement data $\mathbf{y}(\omega) = \mathbf{y}$, one possible estimator that is optimal with respect to many criteria is the random variable $E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]$. Then $(\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)])$ can be interpreted as the random variable to model the error in the estimate: the difference between $\mathbf{x}$ and our estimate of $\mathbf{x}$. The conditional mean of this error vector would be zero. Consequently, $\mathbf{P}_{x|y}$ would be not only the conditional covariance of $\mathbf{x}$,

but also the conditional covariance of the error in our estimate of the value of $\mathbf{x}$.

EXAMPLE 3.15 Let $f_{\mathbf{x}|\mathbf{y}}(\xi|y) = (1/\sqrt{2\pi})\exp\{-\frac{1}{2}(\xi - y)^2\}$. Then the conditional mean and variance of x are

$$E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = y] = y, \qquad E_{\mathbf{x}}\left[(\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = y])^2 \middle| \mathbf{y} = y\right] = 1$$

Thus, for different realizations y of y, the conditional mean is altered but the conditional variance is unchanged for this particular density. ■

To this point, we have assumed the existence of conditional probability density functions. Although this is not a restrictive assumption for our applications, conditional expectations and probabilities can be defined without assuming such existence. Consider a probability space composed of a sample space Ω, σ-algebra $\mathcal{F}$, and probability function P; let $\mathbf{x}$ be a proper random variable (the set $\{\omega:\mathbf{x}(\omega) \leq \xi\}$ is in $\mathcal{F}$). Now let $\mathcal{F}'$ be a σ-algebra that is a subset of $\mathcal{F}$ ($\mathcal{F}'$ is a "coarser" σ-algebra, with fewer or the same number of elements as $\mathcal{F}$). The conditional expectation of $\mathbf{x}$ relative to $\mathcal{F}'$, $E[\mathbf{x}|\mathcal{F}']$, is any ω function that is a proper random variable relative to $\mathcal{F}'$ (the set $\{\omega:E[\mathbf{x}|\mathcal{F}'] \leq \xi\}$ is in $\mathcal{F}'$; i.e., measurable relative to $\mathcal{F}'$) satisfying

$$\int_\Lambda E\{\mathbf{x}|\mathcal{F}'\}\,dP(\omega) = \int_\Lambda \mathbf{x}(\omega)\,dP(\omega) \tag{3-91}$$

for any $\Lambda \subset \Omega$ and $\Lambda \in \mathcal{F}'$. Integration over sets in Ω is defined through measure theory. Letting Λ be Ω itself yields the fact that (3-86) is satisfied by the basic *definition* of a conditional expectation. The *existence* and *uniqueness* of such a random variable is guaranteed by the *Radon–Nikodym theorem* [7, 12, 13, 15], whether or not a density function exists at all.

Now let $\mathbf{y}$ be a vector-valued random variable and let $\mathcal{F}'$ be the minimal σ-algebra with respect to which $\mathbf{y}$ is a proper random variable ($\mathcal{F}'$ is generated by complementation and countable intersection and union of sets of the form $\{\omega:\mathbf{y}(\omega) \leq \rho\}$). Let $B \in \mathcal{F}'$ be the set $\{\omega:\mathbf{y}(\omega) = \mathbf{y}\}$. Then $E\{\mathbf{x}|\mathbf{y} = \mathbf{y}\}$ is defined [7, 15] to be

$$E\{\mathbf{x}|\mathbf{y} = \mathbf{y}\} = E\{\mathbf{x}|\mathcal{F}'\}\Big|_{\omega \in B} \tag{3-92}$$

i.e., it is the random variable function of ω, $E\{\mathbf{x}|\mathcal{F}'\}$, evaluated with a particular ω chosen from the set B.

EXAMPLE 3.16 Let us return to the die-toss experiment described in Examples 3.3, 3.5, 3.8, and 3.11. The payoff random variable $\mathbf{y}(\cdot)$ can be defined through

$$\mathbf{y}(\omega) = \begin{cases} 1 & \text{if } \omega \notin A_1 \text{ or } A_2 \\ 2 & \text{if } \omega \in A_1 \\ 5 & \text{if } \omega \in A_2 \end{cases}$$

with distribution function defined in Fig. 3-15.

Now let z be an indicator of the outcome of the die toss being one of the lower three or upper three numbers:

$$\mathsf{z}(\omega) = \begin{cases} 0 & \text{if } \omega \in \{1 \text{ or } 2 \text{ or } 3 \text{ thrown}\} = A_1 \cup A_2 \\ 1 & \text{if } \omega \in \{4 \text{ or } 5 \text{ or } 6 \text{ thrown}\} = (A_1 \cup A_2)^* \end{cases}$$

The smallest σ-algebra $\mathscr{F}'$ associated with z would be

$$\mathscr{F}' = \{\varnothing, \Omega, A_1 \cup A_2, (A_1 \cup A_2)^*\}$$

which is composed of fewer elements than $\mathscr{F}$ (see Example 3.3). Now let us use (3-91) and (3-63) to generate the conditional expectation, $E\{\mathsf{y}|\mathscr{F}'\}$:

$$\int_\Lambda E\{\mathsf{y}|\mathscr{F}'\}\, dP(\omega) = \int_\Lambda \mathsf{x}(\omega)\, dP(\omega) = \int \rho\, dF_{\mathsf{y}}(\rho) = \sum_i \rho_i\, \Delta F_{\mathsf{y}}(\rho_i)$$

Let $\Lambda = (A_1 \cup A_2) \in \mathscr{F}'$ to obtain

$$\int_{(A_1 \cup A_2)} E\{\mathsf{y}|\mathscr{F}'\}\, dP(\omega) = \sum_i \rho_i\, \Delta F_{\mathsf{y}}(\rho_i) = 2 \cdot \tfrac{1}{3} + 5 \cdot \tfrac{1}{6} = \tfrac{3}{2}$$

But $(A_1 \cup A_2)$ is an atom of $\mathscr{F}'$, so to be a proper random variable on $\mathscr{F}'$, $E\{\mathsf{y}|\mathscr{F}'\}$ must be constant over $(A_1 \cup A_2)$, so this becomes

$$\begin{aligned} \frac{3}{2} &= E\{\mathsf{y}|\mathscr{F}'\}\Big|_{\omega \in (A_1 \cup A_2)} \int_{(A_1 \cup A_2)} dP(\omega) \\ &= E\{\mathsf{y}|\mathscr{F}'\}\Big|_{\omega \in (A_1 \cup A_2)} \left[\frac{1}{3} + \frac{1}{6}\right] \end{aligned}$$

so $E\{\mathsf{y}|\mathscr{F}'\}|_{\omega \in (A_1 \cup A_2)} = 3$. But $\mathsf{z}(\omega) = 0$ if $\omega \in (A_1 \cup A_2)$, so (3-92) yields

$$E\{\mathsf{y}|\mathsf{z} = 0\} = E\{\mathsf{y}|\mathscr{F}'\}\Big|_{\omega \in (A_1 \cup A_2)} = 3.$$

Similar reasoning then yields

$$E\{\mathsf{y}|\mathsf{z} = 1\} = E\{\mathsf{y}|\mathscr{F}'\}\Big|_{\omega \in (A_1 \cup A_2)^*} = 1 \qquad \blacksquare$$

If conditional density functions exist, the definitions in (3-92) and (3-89) are equivalent. We will assume such existence for our applications and will exploit the conditional density function conceptualization.

3.8 CHARACTERISTIC FUNCTIONS

If $\mathbf{x}$ is an n-vector-valued random variable, its *characteristic function* $\phi_{\mathbf{x}}(\cdot)$ is defined as a scalar function of the dummy vector $\boldsymbol{\mu}$ as

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) \triangleq E_{\mathbf{x}}[e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}}] \tag{3-93a}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi}} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\xi_1 \cdots d\xi_n \tag{3-93b}$$

where $j = \sqrt{-1}$. Fourier transform theory can be used to describe the characteristics of $\phi_{\mathbf{x}}$ in terms of the corresponding $f_{\mathbf{x}}$.

One fundamental reason for considering characteristic functions is that *moments* of a random variable can be generated readily through them. Consider

taking the partial derivative of $\phi_{\mathbf{x}}(\boldsymbol{\mu})$ with respect to the kth component of $\boldsymbol{\mu}$, μ_k:

$$\frac{\partial \phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial \mu_k} = j \int_{-\infty}^{\infty} \xi_k e^{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi}} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi}$$

Now divide by j and evaluate the result at $\boldsymbol{\mu} = \mathbf{0}$:

$$\frac{1}{j}\left[\frac{\partial \phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial \mu_k}\right]\Bigg|_{\boldsymbol{\mu}=\mathbf{0}} = \int_{-\infty}^{\infty} \xi_k f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi} = E[\mathsf{x}_k] \tag{3-94}$$

Thus, to obtain the mean of the kth component of $\mathbf{x}$, for $k = 1, 2, \ldots, n$, one can evaluate the partial derivative of $\phi_{\mathbf{x}}(\boldsymbol{\mu})$ with respect to the corresponding component of $\boldsymbol{\mu}$, divide by j, and evaluate the result at $\boldsymbol{\mu} = \mathbf{0}$.

The second moments can be generated through the second partial derivatives. Since

$$\frac{\partial^2 \phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial \mu_k\, \partial \mu_l} = j^2 \int_{-\infty}^{\infty} \xi_k \xi_l e^{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi}} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\boldsymbol{\xi}$$

the second noncentral moment $E[\mathsf{x}_k \mathsf{x}_l]$ can be evaluated as

$$E[\mathsf{x}_k \mathsf{x}_l] = \frac{1}{j^2}\left[\frac{\partial^2 \phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial \mu_k\, \partial \mu_l}\right]\Bigg|_{\boldsymbol{\mu}=\mathbf{0}} \tag{3-95}$$

In general, an Nth noncentral moment of $\mathbf{x}$ can be computed through

$$\underbrace{E[\mathsf{x}_k \mathsf{x}_l \cdots]}_{N \text{ terms}} = \frac{1}{j^N}\left[\frac{\partial^N \phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial \mu_k\, \partial \mu_l \cdots}\right]\Bigg|_{\boldsymbol{\mu}=\mathbf{0}} \tag{3-96}$$

Another application of characteristic functions is the description of the *sum of two independent random variables*. Let $\mathbf{x}$ and $\mathbf{y}$ be two independent n-vector valued random variables, and define $\mathbf{z}$ as their sum,

$$\mathbf{z} = \mathbf{x} + \mathbf{y} \tag{3-97}$$

If we know $f_{\mathbf{x}}(\boldsymbol{\xi})$ and $f_{\mathbf{y}}(\boldsymbol{\rho})$, how can we explicitly generate $f_{\mathbf{z}}(\boldsymbol{\zeta})$? Such a question will arise naturally in estimation when we describe the measurements available to us as true variable values corrupted by additive independent noise. To answer this question, first consider the conditional probability of lying in the infinitesimal hypercube in R^n with one corner at $\boldsymbol{\zeta}$ and of dimension $d\zeta_i = \varepsilon$, $i = 1, 2, \ldots, n$, on each side. Starting with the definition of the conditional density $f_{\mathbf{z}|\mathbf{x}}(\boldsymbol{\zeta}\,|\,\boldsymbol{\xi})$, we can write [see Eq. (3-23)]:

$$\begin{aligned}
f_{\mathbf{z}|\mathbf{x}}(\boldsymbol{\zeta}\,|\,\boldsymbol{\xi})\, d\boldsymbol{\zeta} &= P(\{\omega : \boldsymbol{\zeta} < \mathbf{z}(\omega) \leq \boldsymbol{\zeta} + d\boldsymbol{\zeta}\}, \text{ given that } \mathbf{x}(\omega) = \boldsymbol{\xi}) \\
&= P(\{\omega : \boldsymbol{\zeta} < \mathbf{x}(\omega) + \mathbf{y}(\omega) \leq \boldsymbol{\zeta} + d\boldsymbol{\zeta}\}\,|\,\mathbf{x}(\omega) = \boldsymbol{\xi}) \\
&= P(\{\omega : \boldsymbol{\zeta} - \boldsymbol{\xi} < \mathbf{x}(\omega) + \mathbf{y}(\omega) - \boldsymbol{\xi} \leq \boldsymbol{\zeta} + d\boldsymbol{\zeta} - \boldsymbol{\xi}\}\,|\,\mathbf{x}(\omega) = \boldsymbol{\xi}) \\
&= P(\{\omega : \boldsymbol{\zeta} - \boldsymbol{\xi} < \mathbf{y}(\omega) \leq \boldsymbol{\zeta} - \boldsymbol{\xi} + d\boldsymbol{\zeta}\}\,|\,\mathbf{x}(\omega) = \boldsymbol{\xi})
\end{aligned}$$

But, since $\mathbf{x}$ and $\mathbf{y}$ are independent, this equals the unconditional probability that $\mathbf{y}$ assumes values between the same limits:

$$f_{\mathbf{z}|\mathbf{x}}(\zeta|\xi)\,d\zeta = P(\{\omega:\zeta - \xi < \mathbf{y}(\omega) \le \zeta - \xi + d\zeta\}) = f_{\mathbf{y}}(\zeta - \xi)\,d\zeta$$

Thus, we have shown that

$$f_{\mathbf{z}|\mathbf{x}}(\zeta|\xi) = f_{\mathbf{y}}(\zeta - \xi) \tag{3-98}$$

By combining the concepts of marginal densities and Bayes' rule, this result can be used to write $f_{\mathbf{z}}(\zeta)$ as

$$\begin{aligned} f_{\mathbf{z}}(\zeta) &= \int_{-\infty}^{\infty} f_{\mathbf{z},\mathbf{x}}(\zeta,\xi)\,d\xi = \int_{-\infty}^{\infty} f_{\mathbf{z}|\mathbf{x}}(\zeta|\xi) f_{\mathbf{x}}(\xi)\,d\xi \\ &= \int_{-\infty}^{\infty} f_{\mathbf{y}}(\zeta - \xi) f_{\mathbf{x}}(\xi)\,d\xi \end{aligned} \tag{3-99}$$

This is a *convolution integral*, which is, in general, difficult to evaluate. The corresponding characteristic function is a simple *product*, as expected from the Fourier transform of a convolution.

$$\phi_{\mathbf{z}}(\mu) = E[e^{j\mu^{\mathrm{T}}\mathbf{z}}] = \int_{-\infty}^{\infty} e^{j\mu^{\mathrm{T}}\zeta} f_{\mathbf{z}}(\zeta)\,d\zeta$$

Now substitute (3-99) into this result, letting $\zeta = \xi + \rho$ in the exponential (since $\mathbf{z} = \mathbf{x} + \mathbf{y}$), and assuming we can interchange the order of integration,

$$\begin{aligned} \phi_{\mathbf{z}}(\mu) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\mu^{\mathrm{T}}(\xi+\rho)} f_{\mathbf{y}}(\zeta - \xi) f_{\mathbf{x}}(\xi)\,d\xi \right] d\zeta \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\mu^{\mathrm{T}}(\xi+\rho)} f_{\mathbf{y}}(\zeta - \xi) f_{\mathbf{x}}(\xi)\,d\zeta \right] d\xi \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\mu^{\mathrm{T}}(\xi+\rho)} f_{\mathbf{y}}(\rho) f_{\mathbf{x}}(\xi)\,d\rho \right] d\xi \\ &= \left[\int_{-\infty}^{\infty} e^{j\mu^{\mathrm{T}}\xi} f_{\mathbf{x}}(\xi)\,d\xi \right] \left[\int_{-\infty}^{\infty} e^{j\mu^{\mathrm{T}}\rho} f_{\mathbf{y}}(\rho)\,d\rho \right] \\ &= \phi_{\mathbf{x}}(\mu)\phi_{\mathbf{y}}(\mu) \end{aligned} \tag{3-100}$$

3.9 GAUSSIAN RANDOM VECTORS

A particular random variable of significance to our work is the Gaussian, or normal, vector-valued random variable. First, it provides an adequate model of the random behavior exhibited by many phenomena observed in nature. Second, Gaussian random variables yield tractable mathematical models upon which to base estimators and controllers.

The random n-dimensional vector $\mathbf{x}$ is said to be a *Gaussian (normal) random vector*, or a normally distributed vector-valued random variable, if it can be described through a probability density function of the form

$$f_{\mathbf{x}}(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \exp\{-\tfrac{1}{2}[\boldsymbol{\xi} - \mathbf{m}]^{\mathrm{T}}\mathbf{P}^{-1}[\boldsymbol{\xi} - \mathbf{m}]\} \tag{3-101}$$

where $\mathbf{P}$ is a positive definite $(n \times n)$ matrix, $|\cdot|$ denotes the determinant of a matrix, and $\exp\{\cdot\}$ denotes exponential. The matrix $\mathbf{P}$ must be assumed positive definite to be assured of the existence of $\mathbf{P}^{-1}$. Actually, a more general definition of a Gaussian random vector, allowing positive semidefinite $\mathbf{P}$, can be achieved through the characteristic function. We emphasize the somewhat more restrictive characterization in (3-101) because the density function will provide more physical insight in estimation and control.

Note that the density function in (3-101) is completely defined by the two parameters $\mathbf{m}$ and $\mathbf{P}$. We now claim, and will show later, that these parameters are in fact the mean vector and covariance matrix, respectively. Thus, unlike most other density functions, higher order moments are not required to generate a complete description of the density function.

Figure 3.18 depicts the density function for a scalar Gaussian random variable:

$$f_x(\xi) = \frac{1}{\sqrt{2\pi P}} \exp\left\{-\frac{1}{2P}(\xi - m)^2\right\} \tag{3-102}$$

Because the density is symmetric and unimodal (having one peak), m is both the mean and the mode, the value where the density assumes its peak value. The variance P determines the spread of the density about m as in the figure. If σ is the standard deviation, $\sigma = \sqrt{P}$, then 68.3% of the area under the curve lies in the interval between $(m - \sigma)$ and $(m + \sigma)$. Stated another way, the probability that x assumes a value in the interval $[m - \sigma, m + \sigma]$ is 0.683. Similarly, 95.4% of the probability weight lies between $(m - 2\sigma)$ and $(m + 2\sigma)$, and 99.7% lies between $(m - 3\sigma)$ and $(m + 3\sigma)$. For this reason, peak error specifications are often converted to 3σ values in practice if Gaussian models are employed.

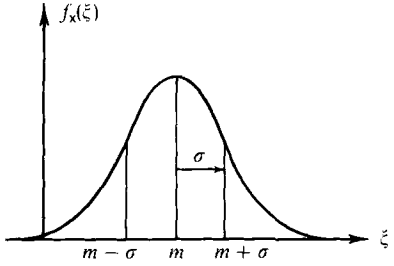

FIG. 3.18 Density function of a scalar Gaussian random variable.

A two-dimensional Gaussian random vector would be characterized by the density function

$$f_{\mathbf{x}}(\boldsymbol{\xi}) = (2\pi)^{-1}\left|\begin{bmatrix}\sigma_1^2 & r_{12}\sigma_1\sigma_2\\ r_{12}\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\right|^{-1/2}$$
$$\times \exp -\frac{1}{2}\left\{\begin{bmatrix}\xi_1 - m_1\\ \xi_2 - m_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\sigma_1^2 & r_{12}\sigma_1\sigma_2\\ r_{12}\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}\xi_1 - m_1\\ \xi_2 - m_2\end{bmatrix}\right\}$$
$$= \frac{1}{2\pi\sigma_1\sigma_2(1 - r_{12}^2)^{1/2}}\exp\left\{-\frac{1}{2(1 - r_{12})^2}\left[\frac{(\xi_1 - m_1)^2}{\sigma_1^2}\right.\right.$$
$$\left.\left. + \frac{(\xi_2 - m_2)^2}{\sigma_2^2} - \frac{2r_{12}(\xi_1 - m_1)(\xi_2 - m_2)}{\sigma_1\sigma_2}\right]\right\} \tag{3-103}$$

This is presented graphically in Fig. 3.19. The mean vector $\mathbf{m}$ in the ξ_1–ξ_2 plane locates the peak of the density function. Loci of constant density function

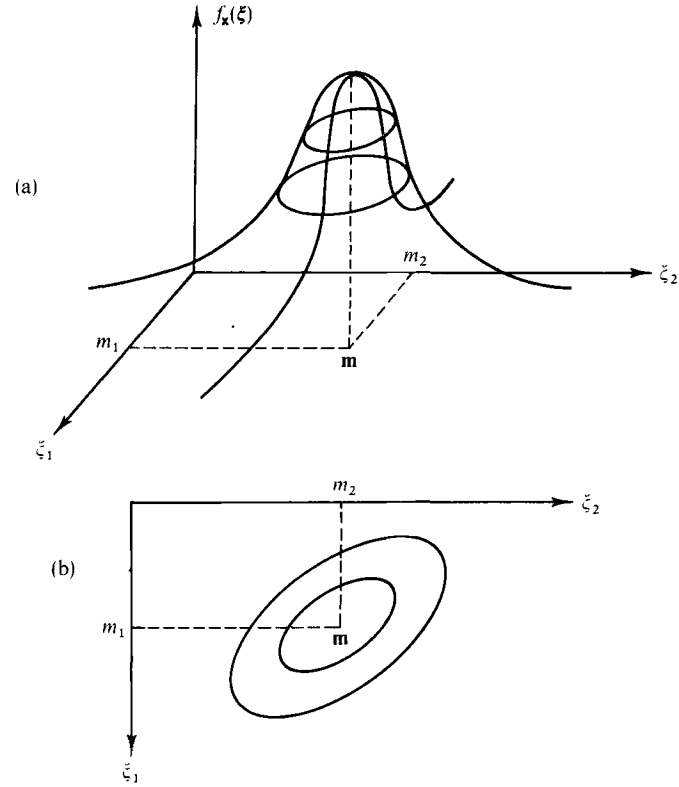

FIG. 3.19 Density function for a two-dimensional Gaussian random vector. (a) Three-dimensional depiction. (b) View from above. The ellipses are the loci of constant probability density value.

values, called surfaces of constant likelihood, are generated by passing planes parallel to the ξ_1–ξ_2 plane through the density function surface, and are ellipses parallel to the ξ_1–ξ_2 plane as shown in the diagram. This can also be seen by setting (3-103) equal to some constant, or equivalently,

$$\frac{(\xi_1 - m_1)^2}{\sigma_1{}^2} + \frac{(\xi_2 - m_2)^2}{\sigma_2{}^2} - \frac{2r_{12}(\xi_1 - m_1)(\xi_2 - m_2)}{\sigma_1\sigma_2} = k \qquad (3\text{-}104)$$

which is the general equation of an ellipse in the ξ_1–ξ_2 plane. Thus the covariance matrix determines the size and angular orientation of ellipses of constant likelihood. If the correlation coefficient r_{12} is zero and thus $\mathbf{P}$ is diagonal, then the principal axes of the ellipses are parallel to the ξ_1 and ξ_2 axes, and $\sigma_1 = \sqrt{P_{11}}$ and $\sigma_2 = \sqrt{P_{22}}$ are the magnitudes of the semimajor and semiminor axes dimensions for the one-sigma ellipse. [This is readily apparent from (3-104).] In general, the eigenvalues of $\mathbf{P}$ provide these magnitudes. If $\mathbf{P}$ is singular and of rank one, then the density function surface collapses down to zero except over a single line (the limit of the ellipses) in the ξ_1–ξ_2 plane: there is no uncertainty in the direction orthogonal to the line since you *know* $\mathbf{x}$ assumes a value somewhere on the line.

These ideas generalize to higher-dimensioned cases, with surfaces of constant likelihood becoming n-dimensional ellipsoids. For example, a probabilistic description of position in three dimensions would be expressible in terms of the size, shape, and orientation of the three-dimensional ellipsoid corresponding to a given probability that the true position lies within the ellipsoid (see Problem 3.11).

The *characteristic function* for a Gaussian random variable $\mathbf{x}$ with density function as in (3-101) is

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) = \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}\boldsymbol{\mu}\} \qquad (3\text{-}105)$$

To show this, we will look at a density function in the general form of (3-101), translate our coordinate system origin to the mean location, and then rotate the coordinates to align them with the principal axes of the ellipsoids of constant likelihood. After working with this simpler description of the density, we will rotate and translate back to the original coordinate system to express the result. The geometric insights gained and linear algebra involved warrant presentation of the details of the proof.

DERIVATION OF CHARACTERISTIC FUNCTION By the definition of characteristic functions, we can write

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi}\} f_{\mathbf{x}}(\boldsymbol{\xi})\, d\xi_1 \cdots d\xi_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi} - \tfrac{1}{2}(\boldsymbol{\xi} - \mathbf{m})^{\mathrm{T}}\mathbf{P}^{-1}(\boldsymbol{\xi} - \mathbf{m})\}\, d\xi_1 \cdots d\xi_n$$

Now translate the coordinate system by making a change of variable $\gamma = \xi - \mathbf{m}$. Note that $d\gamma_i = d\xi_i$ for $i = 1, 2, \ldots, n$. For convenience, define the scalar a as

$$a = \frac{\exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}\}}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}}$$

Thus, $\phi_{\mathbf{x}}(\boldsymbol{\mu})$ becomes

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) = a\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\gamma} - \tfrac{1}{2}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{P}^{-1}\boldsymbol{\gamma}\}\,d\gamma_1\cdots d\gamma_n$$

Having translated the origin of the coordinates to $\mathbf{m}$, rotate the axes into the principal directions. Since $\mathbf{P}$ is symmetric, $\mathbf{P}^{-1}$ is also symmetric. Therefore, there exists an orthogonal transformation matrix $\mathbf{A}$ which diagonalizes $\mathbf{P}^{-1}$: there exists a matrix such that $\mathbf{A}^{\mathrm{T}} = \mathbf{A}^{-1}$ and

$$\mathbf{A}^{\mathrm{T}}\mathbf{P}^{-1}\mathbf{A} = \begin{bmatrix}\sigma_1^{-2} & & 0\\ & \sigma_2^{-2} & \\ & & \ddots\\ 0 & & \sigma_n^{-2}\end{bmatrix} \quad \text{or} \quad \mathbf{P}^{-1} = \mathbf{A}\begin{bmatrix}\sigma_1^{-2} & & 0\\ & \sigma_2^{-2} & \\ & & \ddots\\ 0 & & \sigma_n^{-2}\end{bmatrix}\mathbf{A}^{\mathrm{T}}$$

Thus,

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) = a\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{j\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\gamma} - \tfrac{1}{2}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{A}\begin{bmatrix}\sigma_1^{-2} & & 0\\ & \ddots & \\ 0 & & \sigma_n^{-2}\end{bmatrix}\mathbf{A}^{\mathrm{T}}\boldsymbol{\gamma}\right\}d\gamma_1\cdots d\gamma_n$$

Define a new set of variables through a coordinate rotation as

$$\boldsymbol{\rho} = \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} \leftrightarrow \boldsymbol{\mu} = \mathbf{A}\boldsymbol{\rho}, \qquad \boldsymbol{\zeta} = \mathbf{A}^{\mathrm{T}}\boldsymbol{\gamma} \leftrightarrow \boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\zeta}$$

When we change from integrating over $d\gamma_1\cdots d\gamma_n$ to $d\zeta_1\cdots d\zeta_n$, the Jacobian determinant is equal to one because of the orthogonality of $\mathbf{A}$:

$$d\zeta_1\cdots d\zeta_n = \begin{vmatrix}\partial\zeta_1/\partial\gamma_1 & \cdots & \partial\zeta_1/\partial\gamma_n\\ \vdots & & \vdots\\ \partial\zeta_n/\partial\gamma_1 & \cdots & \partial\zeta_n/\partial\gamma_n\end{vmatrix}d\gamma_1\cdots d\gamma_n = (1)\,d\gamma_1\cdots d\gamma_n$$

With this change of variables, we can write

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) = a\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{j\boldsymbol{\rho}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\zeta} - \tfrac{1}{2}\boldsymbol{\zeta}^{\mathrm{T}}\begin{bmatrix}\sigma_1^{-2} & & 0\\ & \ddots & \\ 0 & & \sigma_n^{-2}\end{bmatrix}\boldsymbol{\zeta}\right\}d\zeta_1\cdots d\zeta_n$$

$$= a\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{\sum_{i=1}^{n}(j\rho_i\zeta_i - \tfrac{1}{2}\zeta_i^2/\sigma_i^2)\right\}d\zeta_1\cdots d\zeta_n$$

$$= a\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^{n}\exp\{j\rho_i\zeta_i - \tfrac{1}{2}\zeta_i^2/\sigma_i^2\}\,d\zeta_1\cdots d\zeta_n$$

$$= a\prod_{i=1}^{n}\int_{-\infty}^{\infty}\exp\{j\rho_i\zeta_i - \tfrac{1}{2}\zeta_i^2/\sigma_i^2\}\,d\zeta_i$$

This is now in the form of a product of n separate one-dimensional integrals. To evaluate each integral, complete the square in the exponential to form

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\{\ \}\, d\zeta_i &= \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}(\zeta_i - j\rho_i\sigma_i^2)^2/\sigma_i^2 - \tfrac{1}{2}\rho_i^2\sigma_i^2\}\, d\zeta_i \\ &= \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}\beta_i^2/\sigma_i^2 - \tfrac{1}{2}\rho_i^2\sigma_i^2\}\, d\beta_i \\ &= \exp\{-\tfrac{1}{2}\rho_i^2\sigma_i^2\} \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}\beta_i^2/\sigma_i^2\}\, d\beta_i\end{aligned}$$

But now the integral term can be recognized as the integral of a scaled Gaussian density of mean zero and variance σ_i^2, so by putting in the proper coefficient, we know the integral equals one:

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\{\ \}\, d\zeta_i &= \sqrt{2\pi}\sigma_i \exp\{-\tfrac{1}{2}\rho_i^2\sigma_i^2\}\left[\frac{1}{\sqrt{2\pi}\sigma_i} \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}\beta_i^2/\sigma_i^2\}\, d\beta_i\right] \\ &= \sqrt{2\pi}\sigma_i \exp\{-\tfrac{1}{2}\rho_i^2\sigma_i^2\}[1]\end{aligned}$$

The characteristic function is now the product of n such integrals as just shown. Writing a explicitly in the expression yields

$$\phi_{\mathbf{x}}(\boldsymbol{\mu}) = \left[\frac{\exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}\}}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}}\right] \prod_{i=1}^{n} (\sqrt{2\pi}\sigma_i \exp\{-\tfrac{1}{2}\rho_i^2\sigma_i^2\})$$

The $(2\pi)^{n/2}$ in the denominator cancels $\prod_{i=1}^{n} \sqrt{2\pi}$. Moreover, since $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ are the eigenvalues of $\mathbf{P}$, $|\mathbf{P}| = \sigma_1^2\sigma_2^2 \cdots \sigma_n^2$ and so $|\mathbf{P}|^{1/2} = \sigma_1\sigma_2 \cdots \sigma_n$, and this then cancels $\prod_{i=1}^{n} \sigma_i$. Thus,

$$\begin{aligned}\phi_{\mathbf{x}}(\boldsymbol{\mu}) &= \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}\} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \rho_i^2\sigma_i^2\right\} \\ &= \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}\} \exp\left\{-\frac{1}{2}\boldsymbol{\rho}^{\mathrm{T}}\begin{bmatrix}\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2\end{bmatrix}\boldsymbol{\rho}\right\} \\ &= \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}\} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\begin{bmatrix}\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2\end{bmatrix}\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}\right\}\end{aligned}$$

Since $\mathbf{P}^{-1} = \mathbf{A}[\sigma_i^{-2}]\mathbf{A}^{\mathrm{T}}$,

$$\mathbf{P} = (\mathbf{A}[\sigma_i^{-2}]\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{\mathrm{T}})^{-1}[\sigma_i^{-2}]^{-1}\mathbf{A}^{-1} = \mathbf{A}[\sigma_i^2]\mathbf{A}^{\mathrm{T}}$$

Substituting this into the quadratic form in $\phi_{\mathbf{x}}(\boldsymbol{\mu})$ yields

$$\begin{aligned}\phi_{\mathbf{x}}(\boldsymbol{\mu}) &= \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}\} \exp\{-\tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}\boldsymbol{\mu}\} \\ &= \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}\boldsymbol{\mu}\}\end{aligned}$$

as claimed in (3-105). ■

Note that the characteristic function does not involve $\mathbf{P}^{-1}$, and therefore it does not inherently require $\mathbf{P}$ to be positive definite. In fact, (3-105) is valid for positive semidefinite $\mathbf{P}$.

The characteristic function can now be used to generate the moments of the Gaussian random vector $\mathbf{x}$. It will be shown now that the parameters $\mathbf{m}$ and

$\mathbf{P}$ in the density and characteristic functions are the *mean* and *covariance* of $\mathbf{x}$, respectively:

$$E[\mathbf{x}] = \mathbf{m} \tag{3-106}$$

$$E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathbf{P} + \mathbf{m}\mathbf{m}^{\mathrm{T}} \tag{3-107}$$

$$E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^{\mathrm{T}}] = \mathbf{P} \tag{3-108}$$

DERIVATION OF MEAN AND COVARIANCE The characteristic function for Gaussian $\mathbf{x}$ is

$$\begin{aligned}\phi_{\mathbf{x}}(\boldsymbol{\mu}) &= \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}\boldsymbol{\mu}\} \\ &= \exp\left\{j\sum_{i=1}^{n}\mu_i m_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{q=1}^{n}\mu_i\mu_q P_{iq}\right\}\end{aligned}$$

To generate the mean of the kth component of $\mathbf{x}$, we take the partial of $\phi_{\mathbf{x}}(\boldsymbol{\mu})$ with respect to μ_k.

$$\frac{\partial\phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial\mu_k} = \left(jm_k - \sum_{q=1}^{n} P_{kq}\mu_q\right)\phi_{\mathbf{x}}(\boldsymbol{\mu})$$

so we can write

$$E[\mathrm{x}_k] = \frac{1}{j}\frac{\partial\phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial\mu_k}\bigg|_{\boldsymbol{\mu}=\mathbf{0}} = m_k$$

Since this is true for all k, $k = 1, 2, \ldots, n$, $E[\mathbf{x}] = \mathbf{m}$.

Using the first partial expression just given, the second partials are

$$\begin{aligned}\frac{\partial^2\phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial\mu_k\,\partial\mu_l} &= \left[\frac{\partial}{\partial\mu_l}\left(jm_k - \sum_{q=1}^{n} P_{kq}\mu_q\right)\right]\phi_{\mathbf{x}}(\boldsymbol{\mu}) + \left(jm_k - \sum_{q=1}^{n} P_{kq}\mu_q\right)\frac{\partial\phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial\mu_l} \\ &= (-P_{kl})\phi_{\mathbf{x}}(\boldsymbol{\mu}) + \left(jm_k - \sum_{q=1}^{n} P_{kq}\mu_q\right)\left(jm_l - \sum_{q=1}^{n} P_{lq}\mu_q\right)\phi_{\mathbf{x}}(\boldsymbol{\mu})\end{aligned}$$

so the second moment $E[\mathrm{x}_k\mathrm{x}_l]$ is

$$E[\mathrm{x}_k\mathrm{x}_l] = \frac{1}{j^2}\frac{\partial^2\phi_{\mathbf{x}}(\boldsymbol{\mu})}{\partial\mu_k\,\partial\mu_l}\bigg|_{\boldsymbol{\mu}=\mathbf{0}} = P_{kl} + m_k m_l$$

This is true for all k and l, so we obtain $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathbf{P} + \mathbf{m}\mathbf{m}^{\mathrm{T}}$. ■

EXAMPLE 3.17 Consider a zero-mean Gaussian random vector, with

$$f_{\mathbf{x}}(\boldsymbol{\xi}) = [(2\pi)^{n/2}|\mathbf{P}|^{1/2}]^{-1}\exp\{-\tfrac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{P}^{-1}\boldsymbol{\xi}\}, \qquad \phi_{\mathbf{x}}(\boldsymbol{\mu}) = \exp\{-\tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}\boldsymbol{\mu}\}$$

For such a random variable, the characteristic function can be used to generate the first four moments as

$$\begin{aligned} E[\mathrm{x}_k] &= 0 & E[\mathrm{x}_k\mathrm{x}_l\mathrm{x}_m] &= 0 \\ E[\mathrm{x}_k\mathrm{x}_l] &= P_{kl} & E[\mathrm{x}_k\mathrm{x}_l\mathrm{x}_m\mathrm{x}_n] &= P_{kl}P_{mn} + P_{km}P_{ln} + P_{kn}P_{lm} \end{aligned}$$ ■

Generalizing the results of the previous example, all odd central moments of a Gaussian random vector are zero (due to symmetry). Moreover, all even central moments can be expressed in terms of the covariance. This is just

another way of saying that the mean and covariance completely define the Gaussian density function.

Previously it was shown that independence implies uncorrelatedness but not necessarily vice versa. It will now be shown that *two jointly Gaussian (normal) random vectors which are uncorrelated are also independent.* We must assume *jointly Gaussian* (defined in the following) random vectors for this to be true—$\mathbf{x}$ and $\mathbf{y}$ can be Gaussian vectors that are *not* jointly Gaussian, and the implication is then not true.

DEMONSTRATION THAT UNCORRELATED → INDEPENDENT IF JOINTLY GAUSSIAN Suppose $\mathbf{x}$ and $\mathbf{y}$ are random vectors, of dimensions n and m, respectively, that are jointly Gaussian and uncorrelated. Define $\mathbf{z}$ to be a random vector of dimension $(n + m)$ composed of components $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{y} \end{bmatrix}$$

To say $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian is equivalent to saying that $\mathbf{z}$ is Gaussian. The first two moments of $\mathbf{z}$ would be

$$\mathbf{m}_z = E[\mathbf{z}] = \begin{bmatrix} E[\mathbf{x}] \\ \hline E[\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{m}_x \\ \hline \mathbf{m}_y \end{bmatrix}$$

$$E[\mathbf{z}\mathbf{z}^T] = \begin{bmatrix} E[\mathbf{x}\mathbf{x}^T] & E[\mathbf{x}\mathbf{y}^T] \\ \hline E[\mathbf{y}\mathbf{x}^T] & E[\mathbf{y}\mathbf{y}^T] \end{bmatrix}; \qquad \mathbf{P}_{zz} = E[\mathbf{z}\mathbf{z}^T] - \mathbf{m}_z\mathbf{m}_z^T$$

Now, since $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated,

$$E[\mathbf{x}\mathbf{y}^T] = \mathbf{m}_x\mathbf{m}_y^T; \qquad E[\mathbf{y}\mathbf{x}^T] = \mathbf{m}_y\mathbf{m}_x^T$$

Thus, the covariance $\mathbf{P}_{zz}$ becomes block diagonal:

$$\mathbf{P}_{zz} = \begin{bmatrix} E[\mathbf{x}\mathbf{x}^T] & \mathbf{m}_x\mathbf{m}_y^T \\ \hline \mathbf{m}_y\mathbf{m}_x^T & E[\mathbf{y}\mathbf{y}^T] \end{bmatrix} - \begin{bmatrix} \mathbf{m}_x\mathbf{m}_x^T & \mathbf{m}_x\mathbf{m}_y^T \\ \hline \mathbf{m}_y\mathbf{m}_x^T & \mathbf{m}_y\mathbf{m}_y^T \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{P}_{yy} \end{bmatrix}$$

Letting ζ be composed of partitions ξ ($\mathbf{x}$ values) and ρ ($\mathbf{y}$ values):

$$\zeta = \begin{bmatrix} \xi \\ \hline \rho \end{bmatrix}$$

the density $f_{\mathbf{z}}(\zeta) = f_{\mathbf{x},\mathbf{y}}(\xi, \rho)$ can now be written as:

$$\begin{aligned}
f_{\mathbf{z}}(\zeta) &= [(2\pi)^{(n+m)/2}|\mathbf{P}_{zz}|^{1/2}]^{-1} \exp\{-\tfrac{1}{2}[\zeta - \mathbf{m}_z]^T\mathbf{P}_{zz}^{-1}[\zeta - \mathbf{m}_z]\} \\
&= \left[(2\pi)^{(n+m)/2}\left|\begin{bmatrix} \mathbf{P}_{xx} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{P}_{yy} \end{bmatrix}\right|^{1/2}\right]^{-1} \exp\left\{-\frac{1}{2}\begin{bmatrix} \xi - \mathbf{m}_x \\ \hline \rho - \mathbf{m}_y \end{bmatrix}^T \begin{bmatrix} \mathbf{P}_{xx}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{P}_{yy}^{-1} \end{bmatrix} \begin{bmatrix} \xi - \mathbf{m}_x \\ \hline \rho - \mathbf{m}_y \end{bmatrix}\right\} \\
&= [(2\pi)^{n/2}(2\pi)^{m/2}|\mathbf{P}_{xx}|^{1/2}|\mathbf{P}_{yy}|^{1/2}]^{-1} \\
&\quad \cdot \exp\{-\tfrac{1}{2}[\xi - \mathbf{m}_x]^T\mathbf{P}_{xx}^{-1}[\xi - \mathbf{m}_x] - \tfrac{1}{2}[\rho - \mathbf{m}_y]^T\mathbf{P}_{yy}^{-1}[\rho - \mathbf{m}_y]\} \\
&= [(2\pi)^{n/2}|\mathbf{P}_{xx}|^{1/2}]^{-1} \exp\{-\tfrac{1}{2}[\xi - \mathbf{m}_x]^T\mathbf{P}_{xx}^{-1}[\xi - \mathbf{m}_x]\} \\
&\quad \cdot [(2\pi)^{m/2}|\mathbf{P}_{yy}|^{1/2}]^{-1} \exp\{-\tfrac{1}{2}[\rho - \mathbf{m}_y]^T\mathbf{P}_{yy}^{-1}[\rho - \mathbf{m}_y]\} \\
&= [f_{\mathbf{x}}(\xi)] \cdot [f_{\mathbf{y}}(\rho)]
\end{aligned}$$

Thus, $\mathbf{x}$ and $\mathbf{y}$ are independent. ■

It was mentioned previously that Gaussian random variables are of engineering importance because they provide adequate models of many random phenomena observed empirically. The basic justification for this statement is embodied in the *central limit theorem*; one of its numerous precise statements (differing in specific assumptions and details, but all essentially the same) is now stated.

CENTRAL LIMIT THEOREM Let $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, be a set of independent random n-vectors which are identically distributed with means and covariance matrices $\mathbf{m}_i$ and $\mathbf{P}_i$, respectively. Define the random vector $\mathbf{y}_N$ as their sum:

$$\mathbf{y}_N = \sum_{i=1}^{N} \mathbf{x}_i$$

and also define $\mathbf{z}_N$ as the (zero-mean) normalized sum random variable:

$$\mathbf{z}_N = [\mathbf{P}_{y_N y_N}]^{-1/2}[\mathbf{y}_N - E[\mathbf{y}_N]]$$

where

$$E[\mathbf{y}_N] = \sum_{i=1}^{N} \mathbf{m}_i, \qquad \mathbf{P}_{y_N y_N} = \sum_{i=1}^{N} \mathbf{P}_i, \qquad \text{and} \qquad \mathbf{P}^{-1/2} = (\mathbf{P}^{1/2})^{-1}$$

where $\mathbf{P}^{1/2}$ is defined as the n-by-n matrix such that $\mathbf{P}^{1/2}\mathbf{P}^{1/2\,\mathrm{T}} = \mathbf{P}$. Then, in the limit as $N \to \infty$, $\mathbf{z}_N$ becomes a zero-mean Gaussian random n-vector with a covariance matrix equal to the identity matrix:

$$\lim_{N \to \infty} f_{\mathbf{z}_N}(\zeta) = [(2\pi)^{n/2}]^{-1} \exp\{-\tfrac{1}{2}\zeta^{\mathrm{T}}\zeta\} \qquad \blacksquare$$

Actually, more general statements can be made, such as not requiring identical distributions for the random variables being summed and then adding some additional, though not very restrictive, assumptions [1–5, 7–11, 14].

Essentially, the theorem states that if the random phenomenon we observe is generated as the sum of effects of many independent random phenomena, then the distribution of the observed phenomenon approaches a Gaussian distribution as more random effects are summed, *regardless* of the distribution of each individual phenomenon. In practice, the assumptions in the theorem are seldom verifiable. Rather, if there are a large number of additive contributing effects to a random phenomenon (as is usually the case when one probes beyond a macroscopic view of a phenomenon), then one suspects that a Gaussian distribution is a reasonable approximation to the actual distribution.

The theorem claims only that a Gaussian distribution is approached as N grows without bound. One would then logically ask, how large does N have to be before the Gaussian approximation is reasonable? The following example due to Papoulis [9] demonstrates a surprisingly good approximation for $N = 3$ and scalar $\mathbf{x}_i$'s uniformly distributed (each thus having a distribution very different from Gaussian).

EXAMPLE 3.18 Let x_1, x_2, and x_3 each be uniformly distributed on the interval $[0, T]$, as in Fig. 3.20a. If $y_2 = x_1 + x_2$, then y_2 has a triangular density function (verifiable by convolution) as in Fig. 3.20b, with mean T and variance $T^2/6$; also plotted is the Gaussian density function with the same first two moments. If $y_3 = x_1 + x_2 + x_3$, its density function consists of three parabolic pieces as in Fig. 3.20c, with mean $3T/2$ and variance $T^2/4$. The normal density with these same statistics is a very good approximation to the true density. ■

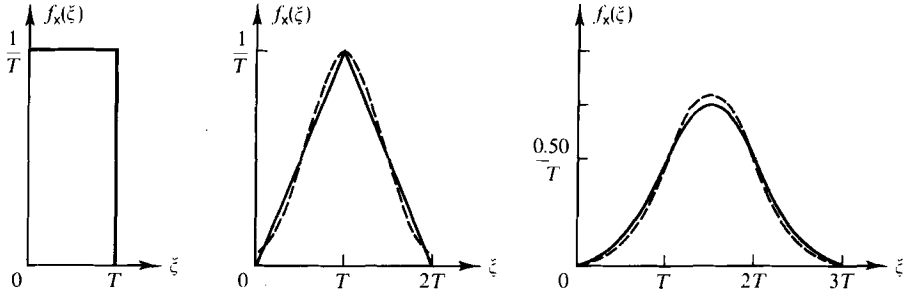

FIG. 3.20 Central limit theorem exemplified. (a) $f_{x_i}(\xi)$. (b) $x = x_1 + x_2$. Solid line indicates $f_x(\xi)$; dashed, $(1/T)\sqrt{3/\pi}\,\exp[-3(x-T)^2/T^2]$. (c) $x = x_1 + x_2 + x_3$. Solid line indicates $f_x(\xi)$; dashed, $(1/T)\sqrt{2/\pi}\,\exp[-2(x-1.5\,T)^2/T^2]$. From *Probability, Random Variables, and Stochastic Processes* by A. Papoulis. © 1965. Used with permission of McGraw-Hill Book Co.

Later when estimation is discussed, the *conditional Gaussian density* will be of primary interest. Therefore, it will be characterized more fully at this point. Let $\mathbf{x}$ and $\mathbf{y}$ be jointly Gaussian vectors mapping Ω into R^n and R^m, respectively, so that $f_{\mathbf{x},\mathbf{y}}(\xi,\rho)$ can be written as

$$f_{\mathbf{x},\mathbf{y}}(\xi,\rho) = \left[(2\pi)^{(n+m)/2}\left|\begin{bmatrix}\mathbf{P}_{xx} & \mathbf{P}_{xy}\\ \mathbf{P}_{yx} & \mathbf{P}_{yy}\end{bmatrix}\right|\right]^{-1} \times \exp\left\{-\frac{1}{2}\begin{bmatrix}\xi-\mathbf{m}_x\\ \rho-\mathbf{m}_y\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{P}_{xx} & \mathbf{P}_{xy}\\ \mathbf{P}_{yx} & \mathbf{P}_{yy}\end{bmatrix}^{-1}\begin{bmatrix}\xi-\mathbf{m}_x\\ \rho-\mathbf{m}_y\end{bmatrix}\right\} \tag{3-109}$$

where we assume that the covariance matrix in (3-109) is positive definite. We claim here, and will prove in the next section, that $\mathbf{x}$ is thus a Gaussian n-vector of mean $\mathbf{m}_x$ and covariance $\mathbf{P}_{xx}$, and $\mathbf{y}$ is a Gaussian m-vector of mean $\mathbf{m}_y$ and covariance $\mathbf{P}_{yy}$. To obtain the conditional density $f_{\mathbf{x}|\mathbf{y}}(\xi\,|\,\rho)$, Bayes' rule can be used to write

$$f_{\mathbf{x}|\mathbf{y}}(\xi\,|\,\rho) = f_{\mathbf{x},\mathbf{y}}(\xi,\rho)/f_{\mathbf{y}}(\rho) \tag{3-110}$$

where $f_{\mathbf{x},\mathbf{y}}(\xi,\rho)$ is given by (3-109) and $f_{\mathbf{y}}(\rho)$ is Gaussian, with moments $\mathbf{m}_y$ and $\mathbf{P}_{yy}$. Performing algebraic reduction yields the result as

$$f_{\mathbf{x}|\mathbf{y}}(\xi\,|\,\rho) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}_{x|y}|^{1/2}}\exp\{-\tfrac{1}{2}[\xi-\mathbf{m}_{x|y}]^{\mathrm{T}}\mathbf{P}_{x|y}^{-1}[\xi-\mathbf{m}_{x|y}]\} \tag{3-111}$$

where

$$\mathbf{m}_{x|y} = \mathbf{m}_x + \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}(\rho-\mathbf{m}_y) \tag{3-112a}$$

$$\mathbf{P}_{x|y} = \mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{yx} \tag{3-112b}$$

Thus, if $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, with joint density given by (3-109), then $f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\xi}|\boldsymbol{\rho})$ is Gaussian with moments $\mathbf{m}_{x|y}$ and $\mathbf{P}_{x|y}$ as just given. The conditional mean of $\mathbf{x}$, given that $\mathbf{y}(\omega) = \mathbf{y}$, is then

$$E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}] \triangleq \mathbf{m}_{x|y} = \mathbf{m}_x + \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}(\mathbf{y} - \mathbf{m}_y) \tag{3-113}$$

From this expression, $E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \cdot]$ can be seen to be an *explicit function* of the realizations $\mathbf{y}$ of $\mathbf{y}$, as stated previously in Section 3.7 for conditional expectations in general. Furthermore, the conditional covariance is:

$$E_{\mathbf{x}}\{[\mathbf{x} - \mathbf{m}_{x|y}][\mathbf{x} - \mathbf{m}_{x|y}]^{\mathrm{T}}|\mathbf{y} = \mathbf{y}\} \triangleq \mathbf{P}_{x|y} = \mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{yx} \tag{3-114}$$

Finally, since $f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\xi}|\boldsymbol{\rho})$ is Gaussian with mean and covariance as described, the conditional characteristic function is:

$$\phi_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\mu}|\boldsymbol{\rho}) = \exp\{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{m}_{x|y} - \tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}_{x|y}\boldsymbol{\mu}\} \tag{3-115}$$

As mentioned before, this function is properly defined for $\mathbf{P}_{x|y}$ positive semi-definite, whereas the density function (3-111) requires $\mathbf{P}_{x|y}$ to be positive definite to ensure the existence of $\mathbf{P}_{x|y}^{-1}$.

If $\mathbf{x}$ represents variables of interest and $\mathbf{y}$ models the measurements available to us, then $f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{\xi}|\boldsymbol{\rho})$ represents the conditional density for the variables of interest, conditioned on knowledge that $\mathbf{y}$ has assumed a particular realization, i.e., conditioned on knowledge of the numerical output of the measuring devices. Should this density be Gaussian, the conditional mean is obviously a valid choice as an estimator for $\mathbf{x}$. (In fact, it will be shown to be an excellent choice under more general conditions as well.) Under such circumstances, the estimator $E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]$ is itself a Gaussian random variable which is a linear combination of the components of $\mathbf{y}(\cdot)$:

$$E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)] = \mathbf{m}_x + \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}[\mathbf{y}(\cdot) - \mathbf{m}_y] \tag{3-116}$$

Moreover, the error in this estimate, $\{\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]\}$, can be shown to be a Gaussian random variable that is *independent* of any random vector obtained by a linear transformation on $\mathbf{y}$: there is no information left in the measurements $\mathbf{y}$ that would yield better insights into the value assumed by $\mathbf{x}$.

3.10 LINEAR OPERATIONS ON GAUSSIAN RANDOM VARIABLES

In developing models for random phenomena and processes, we will want to perform various operations on random variables. If we allow general nonlinear operations on random variables with arbitrary distributions, little can be said in general about the distribution of the transformed variables. However, if we consider linear operations on Gaussian random variables, we can claim that the Gaussian nature is preserved.

First of all, *linear transformations of Gaussian random variables are also Gaussian random variables.* If $\mathbf{x}$ is a Gaussian random n-vector with mean $\mathbf{m}_x$ and covariance $\mathbf{P}_{xx}$, and $\mathbf{A}$ is a known $(m \times n)$ matrix (not random), then the random m-vector $\mathbf{y}$ defined by

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{3-117}$$

is Gaussian with mean and covariance given by

$$\mathbf{m}_y = \mathbf{A}\mathbf{m}_x \tag{3-118a}$$

$$\mathbf{P}_{yy} = \mathbf{A}\mathbf{P}_{xx}\mathbf{A}^{\mathrm{T}} \tag{3-118b}$$

Proof By definition, the characteristic function for $\mathbf{y}$ is

$$\phi_{\mathbf{y}}(\boldsymbol{\mu}) = E[e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{y}}] = E[e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\mathbf{x}}] = E[e^{j[\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}]^{\mathrm{T}}\mathbf{x}}] = \phi_{\mathbf{x}}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu})$$

where the last step is by the definition of $\phi_{\mathbf{x}}(\cdot)$. Since $\mathbf{x}$ is Gaussian, we can write $\phi_{\mathbf{x}}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu})$ explicitly as

$$\phi_{\mathbf{x}}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}) = \exp\{j(\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{m}_x - \tfrac{1}{2}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}_{xx}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu})\} = \exp\{j(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\mathbf{m}_x) - \tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\mathbf{P}_{xx}\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}\}$$

Thus,

$$\phi_{\mathbf{y}}(\boldsymbol{\mu}) = \exp\{j\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{m}_x) - \tfrac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{P}_{xx}\mathbf{A}^{\mathrm{T}})\boldsymbol{\mu}\}$$

which is recognized as the characteristic function of a Gaussian random variable with mean $\mathbf{A}\mathbf{m}_x$ and covariance $\mathbf{A}\mathbf{P}_{xx}\mathbf{A}^{\mathrm{T}}$.

Linear combinations of jointly Gaussian random variables are also Gaussian random variables. Note specifically that we are assuming jointly Gaussian variables here. If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian n- and m-vectors, respectively, and $\mathbf{A}$ and $\mathbf{B}$ are known $(p \times n)$ and $(p \times m)$ matrices, respectively, then the random p-vector $\mathbf{z}$ defined by

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \tag{3-119}$$

is Gaussian, characterized by mean and covariance

$$\mathbf{m}_z = \mathbf{A}\mathbf{m}_x + \mathbf{B}\mathbf{m}_y \tag{3-120a}$$

$$\mathbf{P}_{zz} = \mathbf{A}\mathbf{P}_{xx}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{P}_{xy}\mathbf{B}^{\mathrm{T}} + \mathbf{B}\mathbf{P}_{yx}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{P}_{yy}\mathbf{B}^{\mathrm{T}} \tag{3-120b}$$

Proof Form the $(n + m)$-dimensional Gaussian random variable $\mathbf{w}$

$$\mathbf{w} = \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{y} \end{bmatrix}$$

which is characterized by mean and covariance

$$\mathbf{m}_w = \begin{bmatrix} \mathbf{m}_x \\ \hline \mathbf{m}_y \end{bmatrix}, \qquad \mathbf{P}_{ww} = \left[\begin{array}{c|c} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & \mathbf{P}_{yy} \end{array}\right]$$

Also form the matrix $\mathbf{C} = [\mathbf{A} \mid \mathbf{B}]$. Then $\mathbf{z} = \mathbf{C}\mathbf{w}$, and the result of (3-117) and (3-118) can be invoked. ■

A useful extension of this result is that *linear combinations of jointly Gaussian random variables and nonrandom vectors are also Gaussian random variables.*

If we modify (3-119) to write

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} + \mathbf{c} \tag{3-121}$$

where $\mathbf{c}$ is a known nonrandom p-vector, then $\mathbf{z}$ is a Gaussian random p-vector, with mean and covariance

$$\mathbf{m}_z = \mathbf{A}\mathbf{m}_x + \mathbf{B}\mathbf{m}_y + \mathbf{c} \tag{3-122a}$$

$$\mathbf{P}_{zz} = \mathbf{A}\mathbf{P}_{xx}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{P}_{xy}\mathbf{B}^{\mathrm{T}} + \mathbf{B}\mathbf{P}_{yx}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{P}_{yy}\mathbf{B}^{\mathrm{T}} \tag{3-122b}$$

Proof This is easily proven by following the same steps as in the proof following (3-118), but writing

$$\phi_{\mathbf{z}}(\boldsymbol{\mu}) = E[e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}}] = E[e^{j\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y})}e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{c}}] = e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{c}}E[e^{j\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y})}]$$

The proof is then as before, but with the additional $e^{j\boldsymbol{\mu}^{\mathrm{T}}\mathbf{c}}$ contributing to the mean in $\phi_{\mathbf{z}}(\boldsymbol{\mu})$. ■

Note that only the mean is affected by the addition of $\mathbf{c}$: this makes sense since no uncertainty is contributed by the addition of a known vector. This result will be useful for adding deterministic control inputs to the dynamics model to be developed in the next chapter.

As a final point of interest, the results of (3-117)–(3-118) can be used to show that *any portion of a Gaussian random vector is itself Gaussian*, or equivalently, *if* $\mathbf{x}$ *and* $\mathbf{y}$ *are jointly normal, then their individual marginal densities are also Gaussian.* Let $\mathbf{z}$ be defined as the Gaussian random variable

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{y} \end{bmatrix} \tag{3-123}$$

where $\mathbf{x}$ and $\mathbf{y}$ are n- and m-dimensional partitions, respectively. Assume its mean and covariance are $\mathbf{m}_z$ and $\mathbf{P}_{zz}$, partitioned as

$$\mathbf{m}_z = \begin{bmatrix} \mathbf{m}_x \\ \hline \mathbf{m}_y \end{bmatrix}, \qquad \mathbf{P}_{zz} = \left[\begin{array}{c|c} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & \mathbf{P}_{yy} \end{array}\right] \tag{3-124}$$

Then $\mathbf{x}$ is Gaussian, with mean $\mathbf{m}_x$ and covariance $\mathbf{P}_{xx}$, and $\mathbf{y}$ is Gaussian with first moments $\mathbf{m}_y$ and $\mathbf{P}_{yy}$. This was stated previously [after Eq. (3-109)], but not proven.

Proof The result is obviously true if $\mathbf{x}$ and $\mathbf{y}$ are independent, so that $\mathbf{P}_{xy} = \mathbf{0}$ and $\mathbf{P}_{yx} = \mathbf{0}$. However, it is also true in general, which may not be so obvious. To prove validity in the general case, let $\mathbf{A} = [\mathbf{I}|\mathbf{0}]$ so that

$$\mathbf{x} = \mathbf{A}\mathbf{z} = \mathbf{I}\mathbf{x} + \mathbf{0}\mathbf{y}$$

Then invoke (3-117) through (3-118) to claim $\mathbf{x}$ is a Gaussian random variable with mean and covariance

$$E[\mathbf{x}] = \mathbf{A}\mathbf{m}_z = \mathbf{I}\mathbf{m}_x + \mathbf{0}\mathbf{m}_y = \mathbf{m}_x$$

$$E[(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^{\mathrm{T}}] = \mathbf{A}\mathbf{P}_{zz}\mathbf{A}^{\mathrm{T}} = \mathbf{I}\mathbf{P}_{xx}\mathbf{I}^{\mathrm{T}} + \mathbf{0} = \mathbf{P}_{xx}$$

and similarly for $\mathbf{y}$. ■

3.11 ESTIMATION WITH STATIC LINEAR GAUSSIAN SYSTEM MODELS

A general *estimation problem* can be posed in the following manner. Suppose there are some quantities of interest whose value you do not know exactly. Measuring devices can provide you with data that is functionally related to these variables, but which is also generally noise corrupted. What you would like to do is use this data, and whatever knowledge you have about its relationship to the variables of interest and about its noise corruption, to generate an estimate of the variables under consideration. Furthermore, you would like this estimate to be "optimal" in some sense, where you define the criteria for optimality.

Thus, there are five fundamental components of an estimation problem:

(1) the *variables to be estimated*,

(2) the *measurements* or observations available,

(3) the *mathematical model* describing how the measurements are *related* to the variables of interest,

(4) the *mathematical model* of the *uncertainties* present, and

(5) the *performance evaluation criterion* to judge which estimation algorithms are "best."

We are now able to consider the problem of estimation with static linear Gaussian system models. In other words, we are addressing ourselves to a problem which does not involve dynamics and for which linear system models and Gaussian noise models provide an adequate description. For this case, let us explicitly describe the five problem components just listed.

(1) The variables to be estimated will be put into the form of the components of the n-dimensional vector $\mathbf{x}$. The true values of these quantities will remain constant, but we do not know exactly what the values are.

(2) There will be m measurements available to us, and these will be made the components of an m-dimensional vector, $\mathbf{z}$.

(3) The set of measurement data $\mathbf{z}$ will be assumed to be a linear combination of the variables of interest, corrupted by an uncertain measurement disturbance $\mathbf{v}$ of dimension m:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v} \tag{3-125}$$

where $\mathbf{H}$ is a known $(m \times n)$ matrix.

(4) Probabilistic models will be proposed in the form of random variables to describe the uncertainties (there are other approaches, such as unknown but bounded set descriptions of uncertainties, or "completely unknown" descriptions of disturbances). Thus, our a priori knowledge of the variables of interest can be used in describing the possible values $\mathbf{x}$ as the realizations of a random variable $\mathbf{x}$, assumed to be a Gaussian random variable with mean $\hat{\mathbf{x}}^-$ and

covariance $\mathbf{P}^-$. (The superscript $-$ denotes a value at a time before incorporation of a measurement; $+$ will denote the corresponding value after such incorporation.

Similarly, a random variable model is used to describe the noise corruption. We let $\mathbf{v}$ be a Gaussian random variable, characterized by mean $\mathbf{0}$ and covariance $\mathbf{R}$, and assume that $\mathbf{v}$ and $\mathbf{x}$ are independent. Equation (3-125) can then be viewed as an equation relating the realizations of random variables: for a particular outcome ω, the realization $\mathbf{v}$ of the random variable $\mathbf{v}$ is added to the linear combination $\mathbf{Hx}$ of the realization $\mathbf{x}$ of $\mathbf{x}$ (the particular realization being the "true" value of the variables of interest) to generate the measurement data $\mathbf{z}$. This data $\mathbf{z}$ can itself then be interpreted as the realization of a random variable, denoted as $\mathbf{z}$. Consequently, a random variable model would be

$$\mathbf{z} = \mathbf{Hx} + \mathbf{v} \tag{3-126}$$

where $\mathbf{x}$ and $\mathbf{v}$ are as previously described.

(5) With respect to performance criteria, we will adopt the Bayesian viewpoint that the true objective of our efforts is to generate a complete description of the probability distribution for values of the variables of interest. Since we are interested in estimating the value assumed by a continuous random variable $\mathbf{x}$, knowing the value of the measurement $\mathbf{z}(\omega) = \mathbf{z}$, we are thus really interested in explicitly generating the conditional density function $f_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$. Once such a density function were established, it would provide all the information necessary to define an "optimal" estimate, regardless of the optimality criterion. Consider the general asymmetrical, multipeaked density $f_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$ in Fig. 3.21. Reasonable definitions of an optimal estimate might include the median (having equal probability "weight" on either side), the mode (the peak or maximum likelihood value; hard to distinguish computationally from local peaks), or the mean

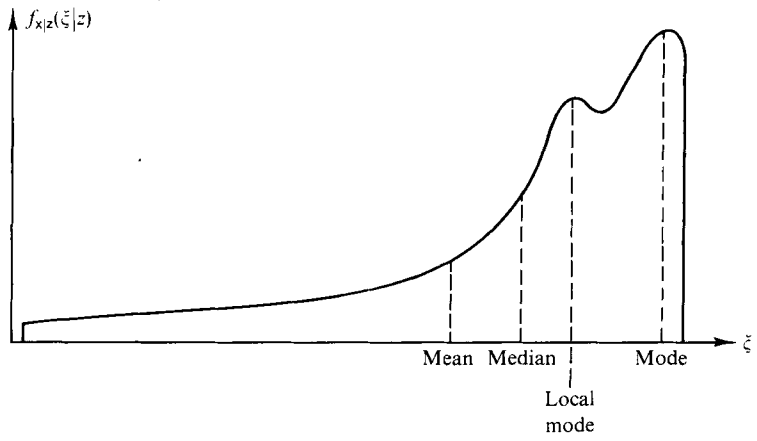

FIG. 3.21 Choice of estimator.

(the "center of probability mass" estimate). By generating the density function, some judgment can be made as to which criterion defines the most reasonable estimate for our purposes, an insight lost by first defining the criterion.

To obtain an explicit evaluation of $f_{\mathbf{x}|\mathbf{z}}(\xi|\mathbf{z})$, we want to show that it is a Gaussian density. In view of the development of (3-109)–(3-113), this entails demonstrating that $\mathbf{x}$ and $\mathbf{z}$ are *jointly* Gaussian random variables. First, since $\mathbf{x}$ and $\mathbf{v}$ are independent Gaussian random variables, they are jointly Gaussian and uncorrelated. This can be shown by reversing the steps taken in Section 3.10 to prove that uncorrelatedness implies independence for jointly Gaussian random variables, writing $f_{\mathbf{x},\mathbf{v}}(\xi, \eta)$ as

$$f_{\mathbf{x},\mathbf{v}}(\xi, \eta) = f_{\mathbf{x}}(\xi) f_{\mathbf{v}}(\eta) \tag{3-127}$$

If we define $\mathbf{u}$ and γ as

$$\mathbf{u} = \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{v} \end{bmatrix}, \qquad \gamma = \begin{bmatrix} \xi \\ \hline \eta \end{bmatrix} \tag{3-128}$$

then $f_{\mathbf{xv}}(\xi, \eta)$ can be written equivalently as $f_{\mathbf{u}}(\gamma)$, a Gaussian density described by mean $\mathbf{m}_u$ and covariance $\mathbf{P}_{uu}$ given by

$$\mathbf{m}_u = \begin{bmatrix} \hat{\mathbf{x}}^- \\ \hline \mathbf{0} \end{bmatrix}, \qquad \mathbf{P}_{uu} = \begin{bmatrix} \mathbf{P}^- & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R} \end{bmatrix} \tag{3-129}$$

So far we have shown $\mathbf{u}$ to be Gaussian. But linear transformations of Gaussian random variables are themselves Gaussian, so $\mathbf{w}$ defined by

$$\mathbf{w} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{H} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{H}\mathbf{x} + \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \hline \mathbf{z} \end{bmatrix} \tag{3-130}$$

is a Gaussian random variable with mean and covariance given by (3-118) as

$$\mathbf{m}_w = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{H} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}^- \\ \hline \mathbf{0} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}^- \\ \hline \mathbf{H}\hat{\mathbf{x}}^- \end{bmatrix} \tag{3-131a}$$

$$\begin{aligned} \mathbf{P}_{ww} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{H} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}^- & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{H}^{\mathrm{T}} \\ \hline \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}^- & \mathbf{P}^-\mathbf{H}^{\mathrm{T}} \\ \hline \mathbf{H}\mathbf{P}^- & \mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R} \end{bmatrix} \end{aligned} \tag{3-131b}$$

Thus, $\mathbf{x}$ and $\mathbf{z}$ are jointly Gaussian random variables, of dimensions n and m, respectively, with their joint density $f_{\mathbf{x},\mathbf{z}}(\xi, \zeta)$ a Gaussian density characterized by $\mathbf{m}_w$ and $\mathbf{P}_{ww}$.

At this point, we can say that $f_{\mathbf{x}|\mathbf{z}}(\xi|\zeta)$ is a Gaussian conditional density function and define it completely through its mean and covariance. Recall the

result of Eqs. (3-109)–(3-113), and make the following replacements:

$$\begin{aligned}
&\text{Random variable:} && \mathbf{y} \to \mathbf{z} \\
&\text{Realization:} && \mathbf{y} \to \mathbf{z} \\
&\text{Dummy variable:} && \boldsymbol{\rho} \to \boldsymbol{\zeta} \\
&\text{Mean of joint density:} && \begin{bmatrix} \mathbf{m}_x \\ \hline \mathbf{m}_y \end{bmatrix} \to \begin{bmatrix} \hat{\mathbf{x}}^- \\ \hline \mathbf{H}\hat{\mathbf{x}}^- \end{bmatrix} \\
&\text{Covariance of joint density:} && \left[\begin{array}{c|c} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & \mathbf{P}_{yy} \end{array}\right] \to \left[\begin{array}{c|c} \mathbf{P}^- & \mathbf{P}^-\mathbf{H}^{\mathrm{T}} \\ \hline \mathbf{H}\mathbf{P}^- & \mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R} \end{array}\right]
\end{aligned}$$

Then the conditional mean, denoted as $\hat{\mathbf{x}}^+$, is given by (3-113) as

$$\hat{\mathbf{x}}^+ = E_{\mathbf{x}}[\mathbf{x}|\mathbf{z} = \mathbf{z}] = \hat{\mathbf{x}}^- + [\mathbf{P}^-\mathbf{H}^{\mathrm{T}}][\mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}^-] \tag{3-132}$$

Similarly (3-114) yields the conditional covariance, denoted by $\mathbf{P}^+$, as

$$\mathbf{P}^+ = \mathbf{P}^- - [\mathbf{P}^-\mathbf{H}^{\mathrm{T}}][\mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{H}\mathbf{P}^-] \tag{3-133}$$

Note that if we define the gain matrix $\mathbf{K}$ as

$$\mathbf{K} = \mathbf{P}^-\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1} \tag{3-134}$$

then (3-132) and (3-133) can be written as

$$\hat{\mathbf{x}}^+ = \hat{\mathbf{x}}^- + \mathbf{K}[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}^-] \tag{3-135}$$

$$\mathbf{P}^+ = \mathbf{P}^- - \mathbf{K}\mathbf{H}\mathbf{P}^- \tag{3-136}$$

Since $\hat{\mathbf{x}}^+$ is the mean of the symmetric Gaussian conditional density $f_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$, it is also the mode. Consequently, we choose it as an optimal estimate of the variables of interest. As discussed at the end of Section 3.5, (3-135) is an equation for a vector $\hat{\mathbf{x}}^+(\mathbf{z})$ in R^n; the mapping $\hat{\mathbf{x}}^+(\cdot)$ from R^m into R^n

$$\hat{\mathbf{x}}^+(\cdot) = \hat{\mathbf{x}}^- + \mathbf{K}[\cdot - \mathbf{H}\hat{\mathbf{x}}^-] \tag{3-137}$$

is an estimator, and the composite mapping $\hat{\mathbf{x}}^+[\mathbf{z}(\cdot)]$ is a random variable

$$\hat{\mathbf{x}}^+ = \hat{\mathbf{x}}^+[\mathbf{z}(\cdot)] = \hat{\mathbf{x}}^- + \mathbf{K}[\mathbf{z}(\cdot) - \mathbf{H}\hat{\mathbf{x}}^-] \tag{3-138}$$

By choosing $\hat{\mathbf{x}}^+$ as an estimate of $\mathbf{x}$, the vector $[\mathbf{x} - \hat{\mathbf{x}}^+]$ is a Gaussian random variable that describes the error in the estimate, denoted as

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}^+ \tag{3-139}$$

The conditional mean of $\mathbf{e}$ is zero, and the conditional covariance (see Problem 3.20) is

$$E_{\mathbf{x}}[\mathbf{e}\mathbf{e}^{\mathrm{T}}|\mathbf{z} = \mathbf{z}] = E_{\mathbf{x}}[(\mathbf{x} - \hat{\mathbf{x}}^+)(\mathbf{x} - \hat{\mathbf{x}}^+)|\mathbf{z} = \mathbf{z}] = \mathbf{P}^+ \tag{3-140}$$

Thus, if we choose $\hat{\mathbf{x}}^+$ as an estimate of $\mathbf{x}$, the $\mathbf{P}^+$ calculated through (3-136) assumes additional significance: it is the covariance to describe the Gaussian error committed by the estimate. Note that this covariance matrix can be computed *without* knowledge of the actual measurement realization, $\mathbf{z}(\omega) = \mathbf{z}$. Consequently, both $\mathbf{P}^+$ and the gain matrix $\mathbf{K}$ can be *precomputed.*

Equations (3-134)–(3-136) can be written in the algebraically equivalent form of

$$\hat{\mathbf{x}}^+ = [\mathbf{P}^+(\mathbf{P}^-)^{-1}]\hat{\mathbf{x}}^- + [\mathbf{P}^+\mathbf{H}^T\mathbf{R}^{-1}]\mathbf{z} \tag{3-141}$$

$$\mathbf{P}^+ = [(\mathbf{P}^-)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1} \tag{3-142}$$

These expressions involve $(n \times n)$ matrix inversions rather than $(m \times m)$ inversions as in the previous equations, so are attractive computationally only if $m > n$; this will be developed further in Sections 5.7 and 7.8. However, this equivalent set of expressions will be more readily manipulated for the case of little or no a priori state information, as will be seen in examples to follow.

EXAMPLE 3.19 Recall the scalar example of the two simultaneous star sightings discussed in Section 1.5. There, x was the one-dimensional position, and $\mathbf{z}$ was the location measured by means of the star sightings, modeled as

$$z_1 = x + v_1, \qquad z_2 = x + v_2$$

We assumed that we had no a priori information about x, that v_1 and v_2 could be modeled as zero-mean Gaussian random variables with variances $\sigma_{z_1}^2$ and $\sigma_{z_2}^2$, respectively, and that x, v_1, and v_2 were independent random variables.

One means of solving for the best estimate of position would be to consider $\mathbf{z}_1(\omega) = z_1$ and the variance $\sigma_{z_1}^2$ to provide the "a priori" information about x before the second measurement is taken. This was the approach taken in Chapter 1: a sequential, or recursive, estimation procedure. Thus, we use this a priori knowledge to describe the random variable x as a Gaussian random variable with mean z_1 and variance $\sigma_{z_1}^2$ ("$\hat{\mathbf{x}}^-$" $= z_1$, "$\mathbf{P}^-$" $= \sigma_{z_1}^2$). Consequently, we consider $\mathbf{z}_2(\omega) = z_2$ as the "available measurement" to be incorporated into the estimate of position. Since we model z_2 as $(x + v_2)$ with v_2 zero-mean, Gaussian, and with variance $\sigma_{z_2}^2$, we have "$\mathbf{R}$" $= \sigma_{z_2}^2$.

The optimal estimate is then the mean (and mode) of the conditional density $f_{x|z_1,z_2}(\xi|z_1,z_2)$:

$$\begin{aligned}\hat{x}^+ &= \hat{x}^- + P^-H^T[HP^-H^T + R]^{-1}(z_2 - H\hat{x}^-) \\ &= z_1 + \sigma_{z_1}^2[\sigma_{z_1}^2 + \sigma_{z_2}^2]^{-1}(z_2 - z_1)\end{aligned}$$

which is, in fact, the result obtained in Eq. (1-6).

The error variance associated with using $\hat{x}^+$ as an estimate, as generated by the estimation algorithm itself, is then P^+:

$$\begin{aligned}P^+ &= P^- - P^-H^T[HP^-H^T + R]^{-1}HP^- \\ &= \sigma_{z_1}^2 - \sigma_{z_1}^2[\sigma_{z_1}^2 + \sigma_{z_2}^2]^{-1}\sigma_{z_1}^2\end{aligned}$$

which was also the result obtained previously, Eq. (1-9). ■

EXAMPLE 3.20 Another means of solving for the best estimate of position in the previous example would be to assume no a priori information about x, and to incorporate the two measurements simultaneously, i.e., in a batch. If there is no a priori information, we could model this through a Gaussian random variable with infinite variance, $P^- = \infty$, or equivalently, $(P^-)^{-1} = 0$.

The measurement is

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} x + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mathbf{H}x + \mathbf{v}$$

where $\mathbf{v}$ is modeled as a zero-mean Gaussian noise of covariance $\mathbf{R}$:

$$\mathbf{R} = E\{\mathbf{v}\mathbf{v}^{\mathrm{T}}\} = \begin{bmatrix} E\{v_1^2\} & E\{v_1 v_2\} \\ E\{v_1 v_2\} & E\{v_2^2\} \end{bmatrix} = \begin{bmatrix} \sigma_{z_1}^2 & 0 \\ 0 & \sigma_{z_2}^2 \end{bmatrix}$$

where the off-diagonal zeros are due to v_1 and v_2 being independent and thus uncorrelated.

Now, using (3-142), we can write P^+ as

$$\begin{aligned} P^+ &= [(P^-)^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}]^{-1} = [\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}]^{-1} \\ &= \left\{ [1 \quad 1] \begin{bmatrix} 1/\sigma_{z_1}^2 & 0 \\ 0 & 1/\sigma_{z_2}^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}^{-1} = \frac{1}{[(1/\sigma_{z_1}^2) + (1/\sigma_{z_2}^2)]} \end{aligned}$$

which is identical to the result of the previous example, and equivalent to the result of Eq. (1-4).

The state estimate can be written using (3-141):

$$\begin{aligned} \hat{x}^+ &= [P^+(P^-)^{-1}]\hat{x}^- + [P^+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}]\mathbf{z} \\ &= P^+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} = [\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} \\ &= \frac{\sigma_{z_1}^2\sigma_{z_2}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} [1 \quad 1] \begin{bmatrix} 1/\sigma_{z_1}^2 & 0 \\ 0 & 1/\sigma_{z_2}^2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \\ &= \frac{\sigma_{z_2}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_1 + \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_2 \end{aligned}$$

Again this is identical to both the previous result and that of Chapter 1. ■

The previous examples demonstrated two methods of processing measurements. In *batch* processing, $\mathbf{z}$ is the vector of all measurements that are available, and thus all measurements are simultaneously incorporated into the estimate. For *recursive* processing, $\mathbf{z}$ is partitioned into components:

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \\ \hline \vdots \\ \hline \mathbf{z}_K \end{bmatrix} \tag{3-143}$$

First, the estimate $\hat{\mathbf{x}}^+$ and covariance $\mathbf{P}^+$ based upon $\mathbf{z}_1(\omega) = \mathbf{z}_1$ alone are computed. Then the Gaussian random variable so obtained, with mean $\hat{\mathbf{x}}^+$ and covariance $\mathbf{P}^+$, is considered to be the information available about $\mathbf{x}$ *prior to the next measurement*, $\mathbf{z}_2(\omega) = \mathbf{z}_2$. The update process is then repeated until all partitions of $\mathbf{z}(\omega) = \mathbf{z}$ are incorporated.

As illustrated by the previous simple examples, if $\mathbf{R}$ is an $(m \times m)$ diagonal matrix, the batch processing of the m-dimensional measurement realization $\mathbf{z}$ and the recursive processing of the m scalar measurements $z_1, z_2, \ldots, z_m$ yield *equivalent* results. To generalize this statement, let $\mathbf{R}$ be block diagonal and

let $\mathbf{z}$ be partitioned corresponding to the diagonal blocks of $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_K \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_K \end{bmatrix} \tag{3-144}$$

Then batch processing of $\mathbf{z}$ and recursive processing of $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$ will yield identical results (see Maybeck [6] and Problem 3.17 for proof). Chapter 7 will extend these results by explicitly generating a transformation of variables to convert any system model into an equivalent form, but with a diagonal $\mathbf{R}$ so that m scalar updates can *always* be used.

This equivalence can be exploited in the design of online estimators. First of all, the recursive form entails the inversion of smaller dimensioned matrices, yielding simpler algorithms. In addition, online estimators are often implemented in general purpose computers, so that only a certain time is allotted to the algorithm, determined in part by the number of high priority interrupts received by the computer to perform other functions. Thus, there *may* not be sufficient time to perform a single batch processing of $\mathbf{z}$ in a given period if the computer is heavily loaded. However, there would be time to process at least *some* of the partitions $\mathbf{z}_1$ to $\mathbf{z}_K$. Since a partially updated estimate would be preferable to one not updated at all, the recursive form might be a substantially better implementation.

The estimation result of Eqs. (3-134)–(3-136) or (3-141)–(3-142) can be directly related to *weighted least squares estimation.* Least squares estimation is a classical technique used extensively, especially in curve fitting applications in which it is desired to obtain the polynomial of given order (or some other chosen functional form) that "best" fits a set of data points. "Best" is defined in terms of minimizing the sum of squares of the differences between the actual measurement data and the proposed, or estimated, function or curve. If one wants to match certain data points more closely than others, a weighting coefficient can be assigned to each term in the sum to be minimized, more heavily weighting the "cost" of differing from the more critical points, yielding what is termed weighted least squares estimation.

Suppose that the measurement model is

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v} \tag{3-145}$$

where $\mathbf{v}$ is an m-vector of measurement noise, whose statistical characteristics are *not* defined. We then want to use our knowledge of the measured value $\mathbf{z}$ to generate an estimate $\hat{\mathbf{x}}$ of the unknown $\mathbf{x}$. Thus, we want to find the value of $\hat{\mathbf{x}}$ that minimizes the weighted sum of squares of the m components of the vector $[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}]$. If we let $\mathbf{W}$ be a general $(m \times m)$ weighting matrix, then we

want to find the vector $\hat{\mathbf{x}}$ that minimizes the scalar cost J:

$$J = \tfrac{1}{2}[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}]^{\mathrm{T}}\mathbf{W}[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}] \tag{3-146}$$

Note that if $\mathbf{W} = \mathbf{I}$, this is standard least squares, with

$$J = \frac{1}{2}\sum_{i=1}^{m}[z - H\hat{x}]_i^2 \tag{3-147}$$

If $\mathbf{W}$ is a diagonal matrix with diagonal terms $w_1, w_2, \ldots, w_m$, then

$$J = \frac{1}{2}\sum_{i=1}^{m} w_i[z - H\hat{x}]_i^2 \tag{3-148}$$

This minimization is accomplished by the value $\hat{\mathbf{x}}_{\mathrm{WLS}}$, where the subscript denotes weighted least squares, if

$$\left.\frac{\partial J}{\partial \hat{\mathbf{x}}}\right|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\mathrm{WLS}}} \triangleq \left.\left[\frac{\partial J}{\partial \hat{x}_1}\ \frac{\partial J}{\partial \hat{x}_2}\cdots\frac{\partial J}{\partial \hat{x}_n}\right]\right|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\mathrm{WLS}}} = \mathbf{0}^{\mathrm{T}} \tag{3-149a}$$

and

$$\left.\frac{\partial^2 J}{\partial \hat{\mathbf{x}}^2}\right|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\mathrm{WLS}}} \geq \mathbf{0} \tag{3-149b}$$

i.e., the second derivative matrix is positive semidefinite. Performing the indicated differentiation on (3-146) yields

$$-[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}]^{\mathrm{T}}\mathbf{W}\mathbf{H}\big|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\mathrm{WLS}}} = \mathbf{0}^{\mathrm{T}}$$

or, $\hat{\mathbf{x}}_{\mathrm{WLS}}$ is the vector that satisfies

$$\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}\hat{\mathbf{x}}_{\mathrm{WLS}} = \mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{z} \tag{3-150}$$

if $[\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}]$ is positive semidefinite. If $[\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}]$ is in fact positive definite, and thus has a unique inverse, then

$$\hat{\mathbf{x}}_{\mathrm{WLS}} = [\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{z} \tag{3-151}$$

This can be compared to the result obtained in Example 3.20 for the case of no a priori information about $\mathbf{x}$, i.e., letting $(\mathbf{P}^-)^{-1} = \mathbf{0}$:

$$\hat{\mathbf{x}}^+ = [\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} \tag{3-152}$$

The two results are identical if we choose $\mathbf{W}$ to be $\mathbf{R}^{-1}$ (positive definite). However, least squares theory gives *no* insights into such a choice of weighting matrix, since no statistical characterization (as $\mathbf{v}$ being Gaussian, zero-mean, of covariance $\mathbf{R}$) was assumed to be known. Analogous to the previous discussion, if $\mathbf{W}$ is block diagonal, then an equivalent result can be achieved by recursive least squares estimation, the form of which is developed in Problem 3.18.

3.12 SUMMARY

This chapter has presented basic concepts of probability theory in a progression that is logical for addressing the problem of estimating some quantities of interest based upon noise-corrupted measurements of related variables. From a Bayesian point of view, this problem is solved by establishing a complete description of the conditional density function of the random vector $\mathbf{x}$ modeling the quantities of interest, conditioned on knowledge of the measured values $\mathbf{z}(\omega_i) = \mathbf{z}$ available: $f_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$. Thus, it was necessary to develop the concepts of random variable models as real-valued functions or mappings, and descriptions of random variables through associated probability distribution and density functions (assuming the existence of the latter). Both unconditioned and conditioned probability functions were discussed, and conditioning allowed the observed realizations of one random variable to provide information about the possible realizations of another, related variable.

The expected value of some function of a random variable is simply the ensemble average value of that function, as the random variable assumes all of its possible realizations. Expectations of particular functions, called moments of a random variable, provided in general a partial description of that random variable. Conditional expectations and moments are of special significance to estimation since it is considerably more feasible to generate and implement algorithms to compute conditional moments than those intended to construct the entire description explicitly as $F_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$ or $f_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$.

In the special case in which $f_{\mathbf{x}|\mathbf{z}}(\boldsymbol{\xi}|\mathbf{z})$ is Gaussian, numerical computation of the first two moments, the mean and covariance, provides a complete depiction of the density function rather than just a partial description. Thus, a computationally feasible estimation algorithm can be developed that satisfies the Bayesian objective of portraying this conditional density. Because linear operations on Gaussian random vectors again yield Gaussian random vectors, the class of problems to which such an algorithm is directly applicable is rather large. This chapter concluded with the detailed development of such an algorithm for estimation with static linear Gaussian system models.

The following chapter will extend these concepts to the case in which quantities of interest can undergo dynamic changes in time. The fundamental ideas of probability theory will be instrumental in developing not only such stochastic process models, but also the estimation and control algorithms that will later exploit these system models.

REFERENCES

1. Bury, K. V., *Statistical Models in Applied Science*. Wiley, New York, 1975.
2. Cramér, H., *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, New Jersey, 1966.
3. Davenport, W. B. Jr., *Probability and Random Processes*. McGraw-Hill, New York, 1970.
4. Feller, W., *An Introduction to Probability and Its Applications*, Vols I and II. Wiley, New York, 1950 (Vol. 1), 1966 (Vol. 2).

5. Loeve, M., *Probability Theory*. Van Nostrand-Reinhold, Princeton, New Jersey, 1963.
6. Maybeck, P. S., "Combined Estimation of States and Parameters for On-Line Applications," Ph.D. dissertation, M.I.T., Cambridge, Massachusetts, February 1972.
7. McGarty, T. P., *Stochastic Systems and State Estimation*. Wiley, New York, 1974.
8. Mendenhall, W., and Schaeffer, R. L., *Mathematical Statistics with Applications*. Duxbury Press, North Scituate, Massachusetts, 1973.
9. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 1965.
10. Parzen, E., *Modern Probability Theory and Its Applications*. Wiley, New York, 1960.
11. Rao, C. R., *Linear Statistical Inference and Its Applications*. Wiley, New York, 1965.
12. Royden, H. L., *Real Analysis*. Macmillan, New York, 1968.
13. Rudin, W., *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1964.
14. Wilks, S. S., *Mathematical Statistics*. Wiley, New York, 1962.
15. Wong, E., *Stochastic Processes in Information and Dynamical Systems*. McGraw-Hill, New York, 1971.

PROBLEMS

3.1 Prove that for any set $A \subset \Omega$, $A \in \mathscr{F}$, the probability $P(A)$ is bounded as $0 \leq P(A) \leq 1$.

3.2 Consider two tosses of a fair coin. Completely define the appropriate probability space $\{\Omega, \mathscr{F}, P\}$. Let us say that we are interested only in the number of heads in the two tosses: can a different probability space $\{\Omega, \mathscr{F}^1, P^1\}$ be defined with $\mathscr{F}^1$ a smaller collection of sets? Define an appropriate random variable $\mathsf{x}(\cdot)$ to consider the number of heads appearing in two tosses. Obtain the probability distribution function for this $\mathsf{x}(\cdot)$.

3.3 If the joint probability density of x_1 and x_2 is

$$f_{\mathsf{x}_1, \mathsf{x}_2}(\xi_1, \xi_2) = \begin{cases} \dfrac{1}{2\pi} \exp[-\tfrac{1}{2}(\xi_1 - 3)^2 - \xi_2] & \xi_2 > 0 \\ 0 & \xi_2 \leq 0 \end{cases}$$

what is the characteristic function for a random variable y, where

$$\mathsf{y} = \mathsf{x}_1 + \mathsf{x}_2 \ ?$$

Determine the mean of y.

3.4 By definition, the ith component of the n-dimensional mean vector is

$$m_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_i f_{\mathsf{x}}(\xi)\, d\xi_1 \cdots d\xi_n$$

Is this the same as

$$m_i = \int_{-\infty}^{\infty} \xi_i f_{\mathsf{x}_i}(\xi_i)\, d\xi_i \ ?$$

Show why.

3.5 Let $\mathsf{x}(\cdot)$ and $\mathsf{y}(\cdot)$ be independent random variables that are each uniformly distributed on the interval $[0, 1]$. Define the random variable $\mathsf{z}(\cdot)$ as

$$\mathsf{z}(\cdot) = \mathsf{x}(\cdot)\mathsf{y}(\cdot)$$

(a) What are the mean, mean squared value, and variance of $\mathsf{x}(\cdot)$ or $\mathsf{y}(\cdot)$? For $\mathsf{z}(\cdot)$?
(b) What is the probability that $\mathsf{z}(\cdot)$ assumes a value less than 0.5? Less than or equal to 0.5?

3.6 Prove that the random vector $\{\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]\}$ is orthogonal to the random vector $\mathbf{y}$: that

$$E_{\mathbf{xy}}\{\mathbf{y}(\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)])^{\mathrm{T}}\} = \mathbf{0}$$

Show that this can be generalized—that $\{\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]\}$ is orthogonal to any function of $\mathbf{y}$. This concept is instrumental in deriving the Kalman filter by means of "orthogonal projections," which was the original means of derivation.

3.7 At the end of Section 3.6, it was stated that if either $\mathbf{x}$ or $\mathbf{y}$ (or both) is zero-mean, then orthogonality and uncorrelatedness of $\mathbf{x}$ and $\mathbf{y}$ imply each other. Prove this. Also prove that, if neither $\mathbf{x}$ nor $\mathbf{y}$ is zero-mean, then they cannot be both uncorrelated and orthogonal.

3.8 Let $\mathbf{x}$ and $\mathbf{y}$ be random n-vectors with $\mathbf{y} = \boldsymbol{\theta}(\mathbf{x})$. Suppose $\boldsymbol{\theta}^{-1}$ exists and that both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{-1}$ are continuously differentiable. Then

$$f_{\mathbf{y}}(\boldsymbol{\rho}) = f_{\mathbf{x}}[\boldsymbol{\theta}^{-1}(\boldsymbol{\rho})]\|\partial\boldsymbol{\theta}^{-1}(\boldsymbol{\rho})/\partial\boldsymbol{\rho}\|$$

where $\|\partial\boldsymbol{\theta}^{-1}(\boldsymbol{\rho})/\partial\boldsymbol{\rho}\| > 0$ is the absolute value of the Jacobian determinant. Prove this theorem using the conditional density relationship

$$f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{\rho}|\boldsymbol{\xi}) = f_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \boldsymbol{\rho})/f_{\mathbf{x}}(\boldsymbol{\xi})$$

as a beginning. Write $f_{\mathbf{y}}(\boldsymbol{\rho})$ in terms of $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi}, \boldsymbol{\rho})$ and continue the proof.

3.9 The scalar Gaussian random variable $\mathsf{x}(\cdot)$ is defined on $\Omega = R^1$ (i.e., the sample space is the real line). The statistics of $\mathsf{x}(\cdot)$ are

$$E\{\mathsf{x}\} = m, \qquad E\{[\mathsf{x} - m]^2\} = P$$

The scalar random variables $\mathsf{y}(\cdot)$ and $\mathsf{z}(\cdot)$ are defined by

$$\mathsf{y}(\cdot) = \mathsf{x}^5(\cdot), \qquad \mathsf{z}(\cdot) = \mathsf{x}^2(\cdot)$$

(a) Find the probability density for $\mathsf{y}(\cdot)$ by using fundamental set definitions to establish $F_{\mathsf{y}}(\rho)$ and then find its derivative.

(b) Find $f_{\mathsf{y}}(\rho)$ by a method analogous to the proof of the preceding problem.

(c) Find $f_{\mathsf{y}}(\rho)$ by direct application of the result of the last problem.

(d) Find the probability density for $\mathsf{z}(\cdot)$ by the method used in part (a). Why would the methods of (b) and (c) not be directly applicable?

3.10 Let x, y, and z be pairwise independent. Show that they need not be triplewise independent. What if $[\mathsf{x}, \mathsf{y}, \mathsf{z}]^{\mathrm{T}}$ is a Gaussian random vector?

3.11 Consider a three-dimensional Gaussian random vector, $\mathbf{x}(\cdot)$, one whose probability density is described by

$$f_{\mathbf{x}}(\boldsymbol{\xi}) = [(2\pi)^{3/2}|\mathbf{P}|^{1/2}]^{-1}\exp\{-\tfrac{1}{2}[\boldsymbol{\xi} - \mathbf{m}]^{\mathrm{T}}\mathbf{P}^{-1}[\boldsymbol{\xi} - \mathbf{m}]\}$$

where the mean $\mathbf{m}$ and covariance $\mathbf{P}$ are

$$m = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 2.5 & 0.5 \\ 0 & 0.5 & 2.5 \end{bmatrix}$$

Surfaces of constant probability density are called surfaces of constant likelihood. They are ellipsoids with principal axes not generally aligned with the coordinate axes.

(a) Determine a transformation of variables $\mathbf{x}' = \mathbf{T}\mathbf{x}$ so that it is possible to use the principal axes of the ellipsoid as the coordinate axes. When this is done, $\mathbf{P}'$ becomes diagonal, i.e.,

$$\mathbf{P}' = \begin{bmatrix} \sigma_{11}'^2 & 0 & 0 \\ 0 & \sigma_{22}'^2 & 0 \\ 0 & 0 & \sigma_{33}'^2 \end{bmatrix}$$

Obtain this form for the given matrix $\mathbf{P}$.

(b) Show that now the surface of constant likelihood is an ellipsoid of the form

$$(\xi_1'^2/\sigma_{11}'^2) + (\xi_2'^2/\sigma_{22}'^2) + (\xi_3'^2/\sigma_{33}'^2) = c^2$$

Write an expression for the probability that x_1, x_2, and x_3 take values within the ellipsoid.

(c) Show that our ellipsoid becomes a sphere by defining new variables

$$\mathsf{x}_1'' = \mathsf{x}_1'/\sigma_{11}', \qquad \mathsf{x}_2'' = \mathsf{x}_2'/\sigma_{22}', \qquad \mathsf{x}_3'' = \mathsf{x}_3'/\sigma_{33}'$$

and that the probability can be written as a volume integral over the ellipsoid:

$$\text{Prob}\{(x_1, x_2, x_3) \text{ lies within ellipsoid}\} = \iiint \frac{e^{-r^2/2}}{(2\pi)^{3/2}} d\xi_1'' \, d\xi_2'' \, d\xi_3''$$

where $r^2 = \xi_1''^2 + \xi_2''^2 + \xi_3''^2$, or in another form as

$$\text{Prob}\{(x_1, x_2, x_3) \text{ lies within ellipsoid}\} = \int_0^c \frac{s(r)e^{-r^2/2}}{(2\pi)^{3/2}} dr$$

where $s(r)$ is the surface area of a sphere of radius r.

(d) Calculate the probability for $c = 1$ and $c = 2$.

3.12 Prove that for a zero-mean Gaussian random vector $\mathbf{x}$, with covariance $\mathbf{P}$,

$$E[\mathsf{x}_k\mathsf{x}_l\mathsf{x}_m\mathsf{x}_n] = P_{kl}P_{mn} + P_{km}P_{ln} + P_{kn}P_{lm}$$

3.13 At the end of Section 3.9, it was claimed that the error $\{\mathbf{x} - E_{\mathbf{x}}[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]\}$ is a Gaussian random vector that is independent of any random vector obtained as a linear transformation on $\mathbf{y}$, under the assumptions made in that section. Prove this.

3.14 A parameter x is to be estimated on the basis of a priori information and a single noisy measurement. The quality of the a priori information is expressed by the probability density function in Fig. 3.P1. The measurement is assumed to be of the form

$$\mathsf{z} = \mathsf{x} + \mathsf{n}$$

where n is a noise, independent of x, which has a probability density of the form given in Fig. 3.P2. The actual measurement taken had the value of $\frac{1}{2}$. Find the conditional probability density $f_{\mathsf{x}|\mathsf{z}}(\xi|\zeta)$ for $\zeta = \frac{1}{2}$, i.e., $f_{\mathsf{x}|\mathsf{z}}(\xi|\frac{1}{2})$. Plot this density as a function of ξ.

One reasonable estimate of x would be the value of ξ that maximizes the density $f_{\mathsf{x}|\mathsf{z}}(\xi|\frac{1}{2})$. This is a "maximum likelihood estimate," and we will denote it here as $\hat{x}_{\text{ML}}$. (It is also called the maximum a posteriori, or MAP, estimate; see Section 5.5.) Find its value.

Another reasonable estimate of x would be the conditional mean, $E_{\mathsf{x}}[\mathsf{x}|\mathsf{z} = \frac{1}{2}]$, which we will denote as $\hat{x}$. Find its value.

Now determine some statistical information about the error committed by these two estimates. Define the error in the maximum likelihood estimate as

$$\mathsf{e}_{\text{ML}} = \hat{\mathsf{x}}_{\text{ML}} - \mathsf{x}$$

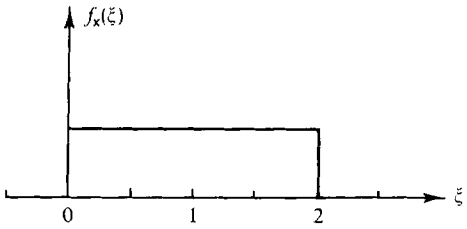

FIG. 3.P1 A priori information for Problem 3.14.

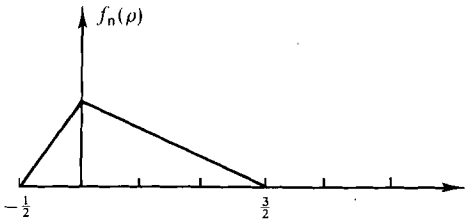

FIG. 3.P2 Measurement noise description for Problem 3.14.

Obtain the conditional mean and conditional variance of this error, conditioned on the fact that $z = \zeta = \frac{1}{2}$. Similarly define the error in the conditional mean estimate of x, and obtain the conditional mean and variance of this error. The conditional mean can be shown to be the estimator, out of the class of linear estimators with zero-mean error, that has minimum error variance: this does *not* mean that other estimators cannot *duplicate* this error variance (as in this problem), or that estimators *outside* of the class under consideration cannot *outperform* the conditional mean.

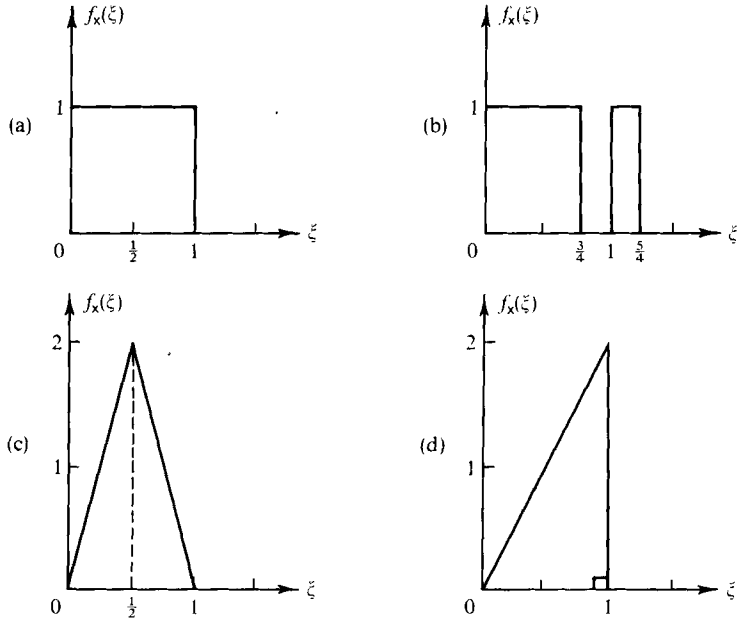

FIG. 3.P3 Density functions for Problem 3.15.

3.15 This problem demonstrates the differences among mean, mode, and median estimates. Find these three estimates for the four density functions depicted in Fig. 3.P3.

3.16 (a) Generate the matrix inverse indicated in Eq. (3-109) by letting

$$\mathbf{P}^{-1} \triangleq \left[\begin{array}{c|c} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & \mathbf{P}_{yy} \end{array}\right]^{-1} \triangleq \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{12}^{\mathrm{T}} & \mathbf{A}_{22} \end{array}\right]$$

and solving $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$ for $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, and $\mathbf{A}_{22}$ as

$$\mathbf{A}_{11} = (\mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{yx})^{-1}, \qquad \mathbf{A}_{22} = (\mathbf{P}_{yy} - \mathbf{P}_{yx}\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy})^{-1}$$
$$\mathbf{A}_{12} = -\mathbf{A}_{11}\mathbf{P}_{xy}\mathbf{P}_{yy}^{-1} = -\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy}\mathbf{A}_{22}$$

(b) Show that equivalent expressions for $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are

$$\mathbf{A}_{11} = \mathbf{P}_{xx}^{-1} + \mathbf{P}_{xx}^{-1}\mathbf{P}_{xy}\mathbf{A}_{22}\mathbf{P}_{yx}\mathbf{P}_{xx}^{-1}, \qquad \mathbf{A}_{22} = \mathbf{P}_{yy}^{-1} + \mathbf{P}_{yy}^{-1}\mathbf{P}_{yx}\mathbf{A}_{11}\mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}$$

Why might this be of use?

(c) Use these results and (3-110) to develop (3-111)–(3-113).

3.17 Prove the claim associated with (3-144) for the case of two measurement vector partitions (an inductive proof for the general case is a simple extension). Let

$$\mathbf{z} = \left[\begin{array}{c} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{H}_1 \\ \hline \mathbf{H}_2 \end{array}\right]\mathbf{x} + \left[\begin{array}{c} \mathbf{v}_1 \\ \hline \mathbf{v}_2 \end{array}\right], \qquad \mathbf{R} = \left[\begin{array}{c|c} \mathbf{R}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_2 \end{array}\right]$$

Show that two recursions of (3-141) and (3-142) for $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively, yields equivalent results to one application of these equations to incorporate $\mathbf{z}$.

3.18 The end of Section 3.11 generated a weighted least squares estimate of $\mathbf{x}$ as given by (3-151):

$$\hat{\mathbf{x}}_{\mathrm{WLS}} = [\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{z}$$

Now convert this into a recursive technique. Let $\mathbf{z}$ be m-dimensional and write the result of the above equation as $\hat{\mathbf{x}}_m$. Now assume an additional scalar measurement value z_{m+1} becomes available; $\hat{\mathbf{x}}_{m+1}$ could be generated in the same manner for an $(m+1)$-dimensional measurement. However, if the "new" values of the $(m+1)$-by-n $\mathbf{H}_{m+1}$ and the $(m+1)$-by-$(m+1)$ $\mathbf{W}_{m+1}$ are

$$\mathbf{H}_{m+1} = \left[\begin{array}{c} \mathbf{H}_m \\ \hline \mathbf{h}_{m+1}^{\mathrm{T}} \end{array}\right], \qquad \mathbf{W}_{m+1} = \left[\begin{array}{c|c} \mathbf{W}_m & \mathbf{0} \\ \hline \mathbf{0} & w_{m+1} \end{array}\right]$$

then show the result can be written equivalently as

$$\tilde{\mathbf{P}}_m = [\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}]^{-1}$$
$$\hat{\mathbf{x}}_{m+1} = \hat{\mathbf{x}}_m + \tilde{\mathbf{P}}_m\mathbf{h}_{m+1}[\mathbf{h}_{m+1}^{\mathrm{T}}\tilde{\mathbf{P}}_m\mathbf{h}_{m+1} + (1/w_{m+1})]^{-1}[z_{m+1} - \mathbf{h}_{m+1}^{\mathrm{T}}\hat{\mathbf{x}}_m]$$
$$\tilde{\mathbf{P}}_{m+1} = \tilde{\mathbf{P}}_m - \tilde{\mathbf{P}}_m\mathbf{h}_{m+1}[\mathbf{h}_{m+1}^{\mathrm{T}}\tilde{\mathbf{P}}_m\mathbf{h}_{m+1} + (1/w_{m+1})]^{-1}\mathbf{h}_{m+1}^{\mathrm{T}}\tilde{\mathbf{P}}_m$$

3.19 Consider the estimation of a vector, $\mathbf{x}$, composed of n constant parameters, based upon m measurements. Let the relationship between the measurements and parameters be given as

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v}$$

Let $\mathbf{x}$ and $\mathbf{v}$ be modeled as jointly Gaussian, zero-mean vectors with

$$E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathbf{X}, \qquad E[\mathbf{v}\mathbf{v}^{\mathrm{T}}] = \mathbf{R}, \qquad E[\mathbf{x}\mathbf{v}^{\mathrm{T}}] = \mathbf{S}$$

What is the conditional mean $E_{\mathbf{x}}[\mathbf{x}|\mathbf{z} = \mathbf{z}]$? If this were used as an estimate of $\mathbf{x}$, what is the corresponding conditional error covariance?

Problem 3.19 is modified from *Uncertain Dynamic Systems* by F. C. Schweppe. © 1973. Used with permission of Prentice-Hall, Inc.

3.20 In Section 3.9 an expression was developed for the conditional covariance of a Gaussian random variable $\mathbf{x}$, conditioned on the fact that a random variable $\mathbf{z}$, jointly Gaussian with $\mathbf{x}$, assumed some value, $\mathbf{z}$. We called this conditional covariance $\mathbf{P}_{x|z}$.

Later, in Section 3.11, this $\mathbf{P}_{x|z}$ was called the covariance of the *error* associated with using $E[\mathbf{x}|\mathbf{z} = \mathbf{z}]$ as an *estimate* of the value of $\mathbf{x}$. Explain this new interpretation of $\mathbf{P}_{x|z}$ explicitly. Show that it is a valid interpretation by showing that if $\mathbf{y}$ is a continuous (Baire) function of $\mathbf{z}$, $\mathbf{y} = \boldsymbol{\theta}(\mathbf{z})$, then

$$E\{\mathbf{x}\mathbf{y}^{\mathrm{T}}|\mathbf{z}(\omega_j) = \mathbf{z}\} = E\{\mathbf{x}|\mathbf{z}(\omega_j) = \mathbf{z}\}[\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{z})]$$

and use this result to prove (3-140).

Is $\mathbf{P}_{x|z}$ a function of the value $\mathbf{z}$ that $\mathbf{z}$ assumes? So what?

3.21 Assume you are responsible for increasing the position-tracking precision of a flight test range. Presently, you have a radar tracking system capable of measuring position ± 10 ft (peak expected errors due to very wideband noise). You desire to double the precision to ± 5 ft approximately. For equal cost you could either

(a) optimally combine the present radar data with a new radar system's data, where the new system alone provides position ± 6 ft (peak errors due to very wide band noise) or

(b) triplicate the original system and combine data optimally.

Which would you propose to do and why? State all modeling assumptions explicitly.

3.22 It is desired to estimate the value assumed by some zero-mean scalar random variable x using the conditional mean of x, given the values of three other zero-mean scalar random variables, z_1, z_2, and z_3. Prove by counterexample that the following two "reasonable" statements are actually false.

(a) If $E[\mathrm{x}\mathrm{z}_1] \neq 0$, then the value of z_1, i.e., z_1, is always a part of the best estimate of $\mathrm{x}(\omega_k) = x$.

(b) If $E[\mathrm{x}\mathrm{z}_1] = 0$, then the value of $\mathrm{z}_1(z_1)$ is never of use in estimating x.

Note what this implies about the common practice in economics and other fields of judging whether a variable should be included in an analysis based on its correlation to the variable of interest.

Problems 3.22 and 3.23 are modified from *Uncertain Dynamic Systems* by F. C. Schweppe. © 1973. Used with permission of Prentice-Hall, Inc.

3.23 The conditional error covariance matrix $\mathbf{P}^{+}$ derived in Section 3.11 is independent of the values $\mathbf{z}$ assumed by the measurements $\mathbf{z}$. Thus, an error analysis can be performed before the measurements are taken, to decide how accurate the estimate will be when (if) it is actually calculated.

Assume that two meters are to provide measurements of a parameter, x, that you want to determine. Let the a priori knowledge of x indicate that it is well modeled as Gaussian, zero-mean, and having a variance of 8. Let the measurements be of the form

$$\mathrm{z}_1 = \mathrm{x} + \mathrm{v}_1, \qquad \mathrm{z}_2 = \mathrm{x} + \mathrm{v}_2$$

One meter has been built and is assumed to be such that v_1 is Gaussian, zero-mean with variance of unity. The other meter has not yet been built, and v_2 can be assumed to be Gaussian, zero-mean, and of variance R, where R is a design parameter.

Assume system specifications require that the final estimate must have an error with variance less than or equal to $\frac{1}{2}$. Since accurate meters cost money, it is reasonable to try to find the maximum value of R that is acceptable. Find this R. With that R, determine the equations for the estimator that incorporates both z_1 and z_2 simultaneously. Now find the equations used to incorporate z_1 to obtain an estimate, and then recursively incorporate z_2 into the estimate. Show that these are the same estimates with the same error variances.

3.24 This and subsequent problems are concerned with estimation of the moments [1, 2, 8, 11, 14] of a random variable $\mathsf{x}(\cdot)$, based only upon N realized values, $x_1, x_2, \ldots, x_N$, that can be considered to be empirical data. The distribution and/or density function are unknown, and assume that the true (unknown) mean and variance of $\mathsf{x}(\cdot)$ are μ_x and σ_x^2, respectively.

A logical choice of estimate of the mean would be

$$\hat{M}_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

To consider the error committed by using this estimate, let $\mathsf{x}_1(\cdot), \mathsf{x}_2(\cdot), \ldots, \mathsf{x}_N(\cdot)$ be N random variables, each with distribution identical to that of $\mathsf{x}(\cdot)$. Thus, we can generate the estimator

$$\hat{\mathsf{M}}_x = \frac{1}{N}\sum_{i=1}^{N} \mathsf{x}_i$$

and conceive of conducting an experiment of generating the N data points repeatedly, the jth such experiment yielding a single realization of $\hat{\mathsf{M}}_x$, $\hat{\mathsf{M}}_x(\omega_j)$. We want to characterize the distribution of these estimate values.

(a) Show that $E\{\hat{\mathsf{M}}_x\} = \mu_x$: that $\hat{\mathsf{M}}_x$ is an unbiased estimator of the mean of $\mathsf{x}(\cdot)$.

(b) Show that the variance of $\hat{\mathsf{M}}_x$, denoted as $\sigma_{\hat{m}}^2$, is

$$\sigma_{\hat{m}}^2 = \frac{1}{N}E\{\mathsf{x}^2\} + \frac{2}{N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E\{\mathsf{x}_i\mathsf{x}_j\} - \mu_x^2$$

so that, if the observations are independent of each other (i.e., $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_N$ are a set of independent random variables), then

$$\sigma_{\hat{m}}^2 = (1/N)\sigma_x^2$$

Why is $\sigma_{\hat{m}}^2$ also the variance of the error committed by using $\hat{M}_x$ to estimate the mean of x?

(c) As N is increased, not only does the estimate become more precise, $\hat{\mathsf{M}}_x$ becomes more and more Gaussian, regardless of the distribution of x. (Why?) If $\hat{\mathsf{M}}_x$ is assumed to be Gaussian, how many observations should be made (i.e., at least how large should N be) so that the probability that the error in the estimate is less than 10% of σ_x is 0.954? (Answer = 400.)

3.25 Since the variance of $\mathsf{x}(\cdot)$ is defined as

$$\sigma_x^2 \triangleq E\{[\mathsf{x} - E\{\mathsf{x}\}]^2\} = E\{\mathsf{x}^2\} - [E\{\mathsf{x}\}]^2$$

a reasonable estimator of the variance of $\mathsf{x}(\cdot)$ would be

$$\hat{V}_x' = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{M}_x)^2 = \frac{1}{N}\left\{\sum_{i=1}^{N} x_i^2\right\} - \hat{M}_x^2$$

with $\hat{M}_x$ defined in the previous problem.

(a) Demonstrate the equality of the two forms of $\hat{V}_x'$.

(b) Although this is a reasonable estimate (and the maximum likelihood estimate), show that it is a biased estimate, that

$$E\{\hat{\mathsf{V}}_x'\} = \sigma_x^2 - \sigma_{\hat{m}}^2 \neq \sigma_x^2$$

with $\sigma_{\hat{m}}^2$ defined in the previous problem. For independent observations, show that this becomes

$$E\{\hat{\mathsf{V}}_x'\} = [(N-1)/N]\sigma_x^2$$

(c) Thus, for independent observations, a variance estimate of $[N/(N-1)]V_x'$, or

$$\hat{V}_x = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{M}_x)^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 - \frac{N}{N-1}\hat{M}_x^2$$

will yield an unbiased estimate of the variance of $\mathsf{x}(\cdot)$, σ_x^2. If observations are not independent, show that

$$E\{\hat{V}_x\} = [N/(N-1)]\{\sigma_x^2 - \sigma_{\hat{m}}^2\}$$

and that the mean error is less than that committed by $\hat{V}_x'$. This estimator is defined only for $N > 1$; would an estimate of variance for $N = 1$ be meaningful? Why would the second form of $\hat{V}_x$ be preferable computationally?

(d) Show that the variance of $\hat{V}_x$, for independent observations and $N > 1$, is

$$\sigma_{\hat{v}}^2 = (1/N)[E\{(\mathsf{x} - \mu_x)^4\} - \{(N-3)/(N-1)\}\sigma_x^4]$$

If $\mathsf{x}(\cdot)$ were assumed Gaussian, show that the fourth central moment is equal to $3\sigma_x^4$, and thus

$$\sigma_{\hat{v}}^2 = [2/(N-1)]\sigma_x^4 \qquad [\text{if } \mathsf{x}(\cdot) \text{ is Gaussian}]$$

If $\mathsf{x}(\cdot)$ were instead assumed to be uniform, show that the fourth central moment is equal to $\frac{9}{5}\sigma_x^4$, and so

$$\sigma_{\hat{v}}^2 = [(4N+6)/\{5N(N-1)\}]\sigma_x^4 \qquad [\text{if } \mathsf{x}(\cdot) \text{ is uniform}]$$

Thus, under very different assumptions about $\mathsf{x}(\cdot)$, the quality of the estimate provided by $\hat{V}_x$ is very similar:

$$\sigma_{\hat{v}} \quad (\mathsf{x} \text{ Gaussian}) \cong 1.4\sigma_{\hat{v}} \quad (\mathsf{x} \text{ uniform})$$

3.26 Analogous to the variance estimator of the previous problem, a good estimate of the covariance between $\mathsf{x}(\cdot)$ and $\mathsf{y}(\cdot)$ is

$$\hat{C}_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{M}_x)(y_i - \hat{M}_y) = \frac{1}{N-1}\sum_{i=1}^{N} x_i y_i - \frac{N}{N-1}\hat{M}_x\hat{M}_y$$

where $\hat{M}_x = (1/N)\sum_{i=1}^{N} x_i$ and $\hat{M}_y = (1/N)\sum_{i=1}^{N} y_i$, and the second form of $\hat{C}_{xy}$ is more convenient computationally.

(a) Show that, for independent observations, $\hat{C}_{xy}$ is an unbiased estimate:

$$E\{\hat{\mathsf{C}}_{xy}\} = \sigma_{xy}^2 = \text{true covariance of } \mathsf{x}(\cdot) \text{ and } \mathsf{y}(\cdot)$$

(b) For independent measurements, the variance of $\hat{\mathsf{C}}_{xy}$ is

$$\sigma_{\hat{c}}^2 = (1/N)[E\{[\mathsf{x} - \mu_x]^2[\mathsf{y} - \mu_y]^2\} + [1/(N-1)]\{\sigma_x^2\sigma_y^2 - (N-2)\sigma_{xy}^4\}]$$

Show that this measure of the quality of the $\hat{C}_{xy}$ estimate becomes, if $\mathsf{x}(\cdot)$ and $\mathsf{y}(\cdot)$ are assumed jointly Gaussian,

$$\sigma_{\hat{c}}^2 = [1/(N-1)]\sigma_x^2\sigma_y^2[1 + r_{xy}^2]$$

where r_{xy} is the correlation coefficient between $\mathsf{x}(\cdot)$ and $\mathsf{y}(\cdot)$.

(c) From the definition of correlation coefficient, (3-75), a good estimator of the (linear) correlation coefficient can be produced as

$$\hat{r}_{xy} = \frac{\hat{C}_{xy}}{(\hat{V}_x\hat{V}_y)^{1/2}} = \frac{N\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{([N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2][N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2])^{1/2}}$$

A "scatter diagram" is a two-dimensional plot of the N realizations $[\mathsf{x}(\omega_1), \mathsf{y}(\omega_1)]$, $[\mathsf{x}(\omega_2)\mathsf{y}(\omega_2)], \ldots,$ $[\mathsf{x}(\omega_N), \mathsf{y}(\omega_N)]$, as shown in Fig. 3.P4. Plot (a) shows perfect correlation between the two variables, (b) portrays smaller positive correlation, (c) depicts no correlation, and (d) shows negative correlation. From plots (a) and (b) it can be seen that the magnitude of the correlation coefficient depicts the dispersion of the points from the least squares (regression) line fit to the data, and not the slope of the line itself. Verify these claims by calculating $\hat{r}_{xy}$ for the four plots of Fig. 3.P4.

The least squares regression line of y on x (least squares fit of a line to the data, with "residuals" being the distance between data points and the line measured in the y direction) is given by

$$y - \hat{M}_y = [\hat{C}_{xy}/\hat{V}_x](x - \hat{M}_x)$$

On the other hand, the least squares regression line of x on y is

$$x - \hat{M}_x = [\hat{C}_{xy}/\hat{V}_y](y - \hat{M}_y)$$

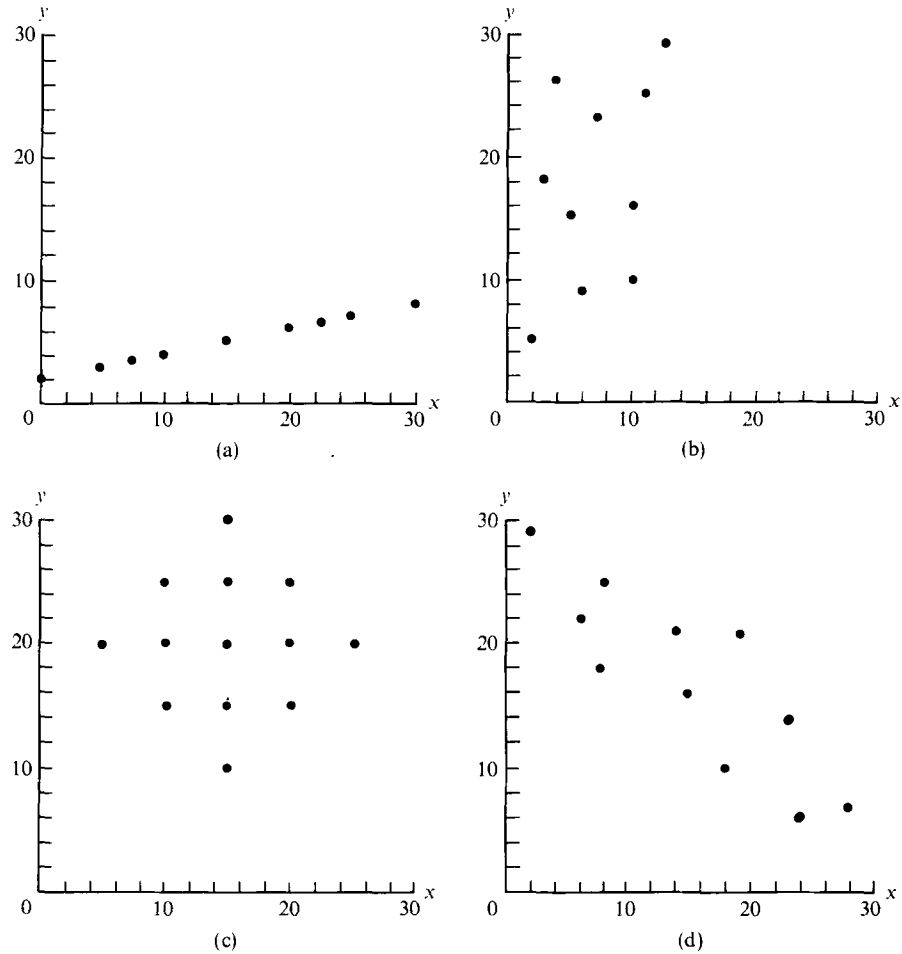

FIG. 3.P4 Scatter diagrams for Problem 3.26.

Thus, each line passes through the "centroid" $(\hat{M}_x, \hat{M}_y)$, and the product of their slopes is $\hat{r}_{xy}^2$. The linear correlation coefficient is a measure of the departure of the two regression lines, with colinearity indicated by $\hat{r}_{xy} = \pm 1$ (perfect linear correlation) and orthogonality by $\hat{r}_{xy} = 0$ (no linear correlation). Verify these interpretations for the four cases depicted in Fig. 3.P4.

3.27 The previous three problems assumed perfect measurements of realized values. Now assume noise corruption of the measuring device, so that what are available are realizations of

$$\mathsf{y}(\cdot) = \mathsf{x}(\cdot) + \mathsf{w}(\cdot)$$

where $\mathsf{w}(\cdot)$ is zero mean, of variance σ_w^2, and independent of $\mathsf{x}(\cdot)$. Assume independent observations.

(a) Show that $\hat{\mathsf{M}}_x = (1/N)\sum_{i=1}^{N} \mathsf{y}_i$ is still an unbiased mean estimator, but with increased variance: $\sigma_{\hat{m}}^2 = (1/N)(\sigma_x^2 + \sigma_w^2)$.

(b) Show that $\hat{\mathsf{V}}_x = [1/(N-1)]\sum_{i=1}^{N} \mathsf{y}_i^2 - [N/(N-1)]\hat{\mathsf{M}}_x^2$, however, is a biased estimator: $E\{\hat{\mathsf{V}}_x\} = \sigma_x^2 + \sigma_w^2$; thus, the best estimator would be $[\hat{\mathsf{V}}_x - \sigma_w^2]$. How would σ_w^2 be estimated? If x and w are assumed Gaussian, show that

$$\sigma_{\hat{v}} \text{ (with meas. noise)} = [1 + (\sigma_w^2/\sigma_x^2)]\sigma_{\hat{v}} \text{ (without meas. noise)}.$$