Stochastic Models, Estimation, and Control Volume 1

Peter S. Maybeck

Stochastic models, estimation, and control

VOLUME 1

Stochastic models, estimation, and control

VOLUME 1

PETER S. MAYBECK

DEPARTMENT OF ELECTRICAL ENGINEERING
AIR FORCE INSTITUTE OF TECHNOLOGY
WRIGHT-PATTERSON AIR FORCE BASE
OHIO

To Beverly

This page intentionally left blank

Contents

Chapter 6   Design and performance analysis of Kalman filters

Chapter 7   Square root filtering

This page intentionally left blank

Preface

The purpose of this book is twofold. First, it attempts to develop a thorough understanding of the *fundamental concepts* incorporated in stochastic processes, estimation, and control. Furthermore, it provides some experience and insights into *applying* the theory to realistic practical problems.

The approach taken is oriented toward an *engineer* or an engineering student. We shall be interested not only in mathematical results, but also in a *physical interpretation* of what the mathematics means. In this regard, considerable effort will be expended to generate graphical representations and to exploit geometric insights where possible. Moreover, our attention will be concentrated upon eventual implementation of estimation and control algorithms, rather than solely upon rigorous derivation of mathematical results in their most general form. For example, all assumptions will be described thoroughly in order to yield precise results, but these assumptions will further be *justified* and their *practical implications* pointed out explicitly. Topics where additional generality or rigor can be incorporated will also be delineated, so that such advanced disciplines as functional analysis can be exploited by, but are not required of, the reader.

Because this book is written for engineers, we shall avoid measure theory, functional analysis, and other disciplines that may not be in an engineer's background. Although these fields can make the theoretical developments more rigorous and complete, they are not essential to the practicing engineer who wants to *use* optimal estimation theory results. Furthermore, the book can serve as a text for a first-year graduate course in estimation and stochastic control, and these advanced disciplines are not generally studied prior to such a course. However, the places where these disciplines do contribute will be pointed out for those interested in pursuing more rigorous developments. The developments in the text will also be motivated in part by the concepts of analysis and functional analysis, but without requiring the reader

to be previously familiar with these fields. In this way, the reader will become aware of the kinds of questions that have to be answered in a completely rigorous derivation and will be introduced to the concepts required to resolve them properly.

This work is intended to be a text from which a reader can *learn* about estimation and stochastic control, and this intent has dictated a format of presentation. Rather than strive for the mathematical precision of a theorem–proof structure, fundamentals are first motivated conceptually and physically, and then the mathematics developed to serve the purpose. Practical aspects and eventual implementation of algorithms are kept at the forefront of concern. Finally, the progression of topics is selected to maximize learning: a firm foundation in linear system applications is laid before nonlinear applications are considered, conditional probability density functions are discussed before conditional expectations, and so forth. Although a reference book might be organized from the most general concepts progressively to simpler and simpler special cases, it has been our experience that people grasp basic ideas and understand complexities of the general case better if they build up from the simpler problems. As generalizations are made in the text, care is taken to point out all ramifications—what changes are made in the previous simpler case, what concepts generalize and how, what concepts no longer apply, and so forth.

With an eye to practicality and eventual implementations, we shall emphasize the case of continuous-time dynamic systems with *discrete-time* data sampling. Most applications will in fact concern continuous-time systems, while the actual estimator or controller implementations will be in the form of software algorithms for a digital computer, which inherently involves data samples. These *algorithms* will be developed in detail, with special emphasis on the various *design tradeoffs* involved in achieving an efficient, practical configuration.

The corresponding results for the case of continuously available measurements will be presented, and its range of applicability discussed. However, only a formal derivation of the results will be provided; a rigorous derivation, though mathematically enriching, does not seem warranted because of this limited applicability. Rather, we shall try to develop physical insights and an engineering appreciation for these results.

Throughout the development, we shall regard the digital computer not only as the means for eventual implementation of on-line algorithms, but also as a *design tool* for generating the final "tuned" algorithms themselves. We shall develop means of synthesizing estimators or controllers, fully evaluating their performance capabilities in a real-world environment, and iterating upon the design until performance is as desired, all facilitated by software tools.

Because the orientation is toward engineering applications, *examples* will

be exploited whenever possible. Unfortunately, even under our early restrictions of a linear system model driven by white Gaussian noises (these assumptions will be explained later), simple estimation or control examples are difficult to generate—either they are simple enough to work manually and are of little value, or are useful, enlightening, and virtually impossible to do by hand. At first, we shall try to gain *insights* into algorithm structure and behavior by solving relatively simple problems. Later, more complex and realistic problems will be considered in order to appreciate the *practical aspects* of estimator or controller implementation.

This book is the outgrowth of the first course of a two-quarter sequence taught at the Air Force Institute of Technology. Students had previously taken a course in applied probability theory, taught from the excellent Chapters 1–7 of Davenport's "Probability and Random Processes." Many had also been exposed to a first control theory course, linear algebra, linear system theory, deterministic optimal control, and random processes. However, considerable attention is paid to those fundamentals in Chapters 2–4, before estimation and stochastic control are developed at all. This has been done out of the conviction that system modeling is a critical aspect, and typically the "weak link," in applying theory to practice.

Thus the book has been designed to be self-contained. The reader is assumed to have been exposed to advanced calculus, differential equations, and some vector and matrix analysis on an engineering level. Any more advanced mathematical concepts will be developed within the text, requiring only a willingness on the part of the reader to deal with new means of conceiving a problem and its solution. Although the mathematics becomes relatively sophisticated at times, efforts are made to motivate the need for, and stress the underlying basis of, this sophistication. The objective is to investigate the theory and derive from it the tools required to reach the ultimate objective of generating practical designs for estimators and stochastic controllers.

The author wishes to express his gratitude to the students who have contributed significantly to the writing of this book through their helpful suggestions and encouragement. The stimulation of technical discussions and association with Professors John Deyst, Wallace Vander Velde, and William Widnall of the Massachusetts Institute of Technology and Professors Jurgen Gobien, James Negro, and J. B. Peterson of the Air Force Institute of Technology has also had a profound effect on this work. Appreciation is expressed to Dr. Robert Fontana, Head of the Department of Electrical Engineering, Air Force Institute of Technology, for his support throughout this endeavor, and to those who carefully and thoroughly reviewed the manuscript. I am also deeply grateful to my wife, Beverly, whose understanding and constant supportiveness made the fruition of this effort possible.

This page intentionally left blank

Contents of Volume 2

This page intentionally left blank

Notation

Vectors, Matrices

    Scalars are denoted by upper or lower case letters in italic type.

    Vectors are denoted by lower case letters in boldface type, as the vector **x** made up of components x_i.

    Matrices are denoted by upper case letters in boldface type, as the matrix **A** made up of elements A_{ij} (*i*th row, *j*th column).

Random Vectors (Stochastic Processes), Realizations (Samples),
and Dummy Variables

    Random vectors are set in boldface sans serif type, as **x** made up of scalar components x_i.

    Realizations of the random vector are set in boldface roman type, as **x**: $\mathbf{x}(\omega_i) = \mathbf{x}$.

    Dummy variables (for arguments of density or distribution functions, integrations, etc.) are denoted by the equivalent Greek letter, such as ξ being associated with **x**: e.g., $f_{\mathsf{x}}(\xi)$. The correspondences are $(\mathsf{x}, \xi), (\mathsf{y}, \rho), (\mathsf{z}, \zeta), (\mathsf{Z}, \mathscr{Z})$.

Subscripts

    a: augmented   c: continuous-time
    d: discrete-time   t: true, truth model

Superscripts

    ^{T}: transpose (matrix) $^-$: Fourier transform
    $^{-1}$: inverse (matrix) $^{\wedge}$: estimate
    *: complement (set) or complex conjugate

Matrix and Vector Relationships

    $\mathbf{A} > \mathbf{0}$: **A** is positive definite.
    $\mathbf{A} \geq \mathbf{0}$: **A** is positive semidefinite.
    $\mathbf{x} \leq \mathbf{a}$: componentwise, $x_1 \leq a_1, x_2 \leq a_2, \ldots,$ and $x_n \leq a_n$.

List of symbols and pages where they are defined or first used

ϕ_x	99	$\Psi_{xx}(\tau)$	140
Ψ_{xx}	90	$\bar{\Psi}_{xx}(\omega)$	141
$\Psi_{xx}(t)$	137	Ω	60
$\Psi_{xx}(t, t + \tau)$	137	ω	60

This page intentionally left blank

Introduction

1.1 WHY STOCHASTIC MODELS, ESTIMATION, AND CONTROL?

When considering system analysis or controller design, the engineer has at his disposal a wealth of knowledge derived from *deterministic* system and control theories. One would then naturally ask, why do we have to go beyond these results and propose *stochastic* system models, with ensuing concepts of estimation and control based upon these stochastic models? To answer this question, let us examine what the deterministic theories provide and determine where the shortcomings might be.

Given a physical system, whether it be an aircraft, a chemical process, or the national economy, an engineer first attempts to develop a mathematical model that adequately represents some aspects of the behavior of that system. Through physical insights, fundamental "laws," and empirical testing, he tries to establish the interrelationships among certain variables of interest, inputs to the system, and outputs from the system.

With such a mathematical model and the tools provided by system and control theories, he is able to investigate the system structure and modes of response. If desired, he can design compensators that alter these characteristics and controllers that provide appropriate inputs to generate desired system responses.

In order to observe the actual system behavior, measurement devices are constructed to output data signals proportional to certain variables of interest. These output signals and the known inputs to the system are the only information that is directly discernible about the system behavior. Moreover, if a feedback controller is being designed, the measurement device outputs are the only signals directly available for inputs to the controller.

There are three basic reasons why deterministic system and control theories do not provide a totally sufficient means of performing this analysis and

design. First of all, *no mathematical system model is perfect.* Any such model depicts only those characteristics of direct interest to the engineer's purpose. For instance, although an endless number of bending modes would be required to depict vehicle bending precisely, only a finite number of modes would be included in a useful model. The objective of the model is to represent the dominant or critical modes of system response, so many effects are knowingly left unmodeled. In fact, models used for generating online data processors or controllers must be pared to only the basic essentials in order to generate a computationally feasible algorithm.

Even effects which are modeled are necessarily *approximated* by a mathematical model. The "laws" of Newtonian physics are adequate approximations to what is actually observed, partially due to our being unaccustomed to speeds near that of light. It is often the case that such "laws" provide adequate system *structures,* but various *parameters* within that structure are not determined absolutely. Thus, there are many sources of uncertainty in any mathematical model of a system.

A second shortcoming of deterministic models is that dynamic systems are driven not only by our own control inputs, but also by *disturbances which we can neither control nor model deterministically.* If a pilot tries to command a certain angular orientation of his aircraft, the actual response will differ from his expectation due to wind buffeting, imprecision of control surface actuator responses, and even his inability to generate exactly the desired response from his own arms and hands on the control stick.

A final shortcoming is that *sensors do not provide perfect and complete data* about a system. First, they generally do not provide all the information we would like to know: either a device cannot be devised to generate a measurement of a desired variable or the cost (volume, weight, monetary, etc.) of including such a measurement is prohibitive. In other situations, a number of different devices yield functionally related signals, and one must then ask how to generate a best estimate of the variables of interest based on partially redundant data. Sensors do not provide exact readings of desired quantities, but introduce their own system dynamics and distortions as well. Furthermore, these devices are also always noise corrupted.

As can be seen from the preceding discussion, to assume perfect knowledge of all quantities necessary to describe a system completely and/or to assume perfect control over the system is a naive, and often inadequate, approach. This motivates us to ask the following four questions:

(1) How do you develop system models that account for these uncertainties in a direct and proper, yet practical, fashion?

(2) Equipped with such models and incomplete, noise-corrupted data from available sensors, how do you optimally estimate the quantities of interest to you?

(3)   In the face of uncertain system descriptions, incomplete and noise-corrupted data, and disturbances beyond your control, how do you optimally control a system to perform in a desirable manner?

(4)   How do you evaluate the performance capabilities of such estimation and control systems, both before and after they are actually built?

This book has been organized specifically to answer these questions in a meaningful and useful manner.

1.2   OVERVIEW OF THE TEXT

Chapters 2–4 are devoted to the stochastic modeling problem. First Chapter 2 reviews the pertinent aspects of deterministic system models, to be exploited and generalized subsequently. Probability theory provides the basis of all of our stochastic models, and Chapter 3 develops both the general concepts and the natural result of static system models. In order to incorporate dynamics into the model, Chapter 4 investigates stochastic processes, concluding with practical linear dynamic system models. The basic form is a linear system driven by white Gaussian noise, from which are available linear measurements which are similarly corrupted by white Gaussian noise. This structure is justified extensively, and means of describing a large class of problems in this context are delineated.

Linear estimation is the subject of the remaining chapters. Optimal filtering for cases in which a linear system model adequately describes the problem dynamics is studied in Chapter 5. With this background, Chapter 6 describes the design and performance analysis of practical online Kalman filters. Square root filters have emerged as a means of solving some numerical precision difficulties encountered when optimal filters are implemented on restricted wordlength online computers, and these are detailed in Chapter 7.

Volume 1 is a complete text in and of itself. Nevertheless, Volume 2 will extend the concepts of linear estimation to smoothing, compensation of model inadequacies, system identification, and adaptive filtering. Nonlinear stochastic system models and estimators based upon them will then be fully developed. Finally, the theory and practical design of stochastic controllers will be described.

1.3   THE KALMAN FILTER:
AN INTRODUCTION TO CONCEPTS

Before we delve into the details of the text, it would be useful to see where we are going on a conceptual basis. Therefore, the rest of this chapter will provide an overview of the optimal linear estimator, the Kalman filter. This will be conducted at a very elementary level but will provide insights into the

underlying concepts. As we progress through this overview, contemplate the ideas being presented: try to conceive of graphic *images* to portray the concepts involved (such as time propagation of density functions), and to generate a *logical structure* for the component pieces that are brought together to solve the estimation problem. If this basic conceptual framework makes sense to you, then you will better understand the need for the details to be developed later in the text. Should the idea of where we are going ever become blurred by the development of detail, refer back to this overview to regain sight of the overall objectives.

First one must ask, what is a Kalman filter? A Kalman filter is simply an *optimal recursive data processing algorithm*. There are many ways of defining *optimal*, dependent upon the criteria chosen to evaluate performance. It will be shown that, under the assumptions to be made in the next section, the Kalman filter is optimal with respect to virtually any criterion that makes sense. One aspect of this optimality is that the Kalman filter incorporates all information that can be provided to it. It processes all available measurements, regardless of their precision, to estimate the current value of the variables of interest, with use of (1) knowledge of the system and measurement device dynamics, (2) the statistical description of the system noises, measurement errors, and uncertainty in the dynamics models, and (3) any available information about initial conditions of the variables of interest. For example, to determine the velocity of an aircraft, one could use a Doppler radar, or the velocity indications of an inertial navigation system, or the pitot and static pressure and relative wind information in the air data system. Rather than ignore any of these outputs, a Kalman filter could be built to combine all of this data and knowledge of the various systems' dynamics to generate an overall best estimate of velocity.

The word *recursive* in the previous description means that, unlike certain data processing concepts, the Kalman filter does not require all previous data to be kept in storage and reprocessed every time a new measurement is taken. This will be of vital importance to the practicality of filter implementation.

The "filter" is actually a *data processing algorithm*. Despite the typical connotation of a filter as a "black box" containing electrical networks, the fact is that in most practical applications, the "filter" is just a computer program in a central processor. As such, it inherently incorporates discrete-time measurement samples rather than continuous time inputs.

Figure 1.1 depicts a typical situation in which a Kalman filter could be used advantageously. A system of some sort is driven by some known controls, and measuring devices provide the value of certain pertinent quantities. Knowledge of these system inputs and outputs is all that is explicitly available from the physical system for estimation purposes.

The *need* for a filter now becomes apparent. Often the variables of interest, some finite number of quantities to describe the "state" of the system, cannot
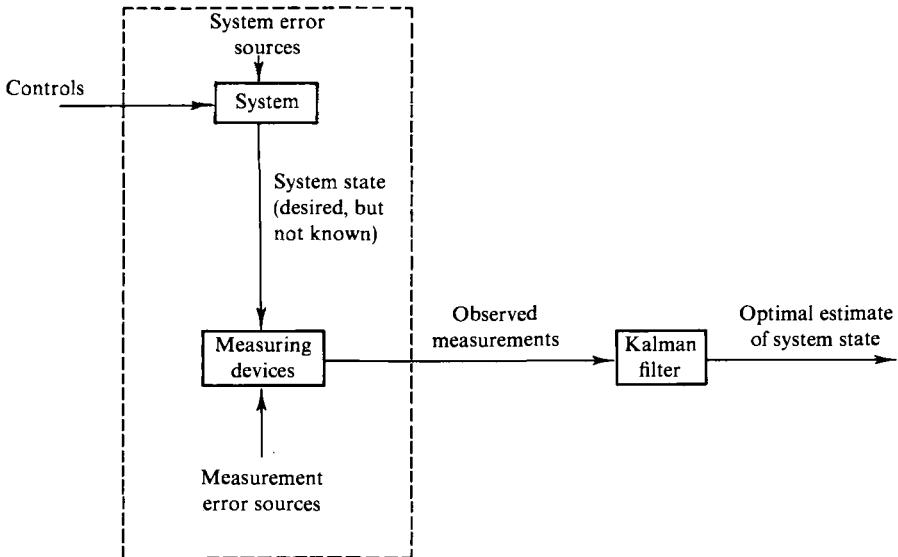
FIG. 1.1   Typical Kalman filter application.

be measured directly, and some means of inferring these values from the avail-
able data must be generated. For instance, an air data system directly provides
static and pitot pressures, from which velocity must be inferred. This inference
is complicated by the facts that the system is typically driven by inputs other
than our own known controls and that the relationships among the various
"state" variables and measured outputs are known only with some degree of
uncertainty. Furthermore, any measurement will be corrupted to some degree
by noise, biases, and device inaccuracies, and so a means of extracting valuable
information from a noisy signal must be provided as well. There may also be
a number of different measuring devices, each with its own particular dynamics
and error characteristics, that provide some information about a particular
variable, and it would be desirable to combine their outputs in a systematic
and optimal manner. A Kalman filter combines all available measurement data,
plus prior knowledge about the system and measuring devices, to produce an
estimate of the desired variables in such a manner that the error is minimized
statistically. In other words, if we were to run a number of candidate filters
many times for the same application, then the average results of the Kalman
filter would be better than the average results of any other.

Conceptually, what any type of filter tries to do is obtain an "optimal"
estimate of desired quantities from data provided by a noisy environment,
"optimal" meaning that it minimizes errors in some respect. There are many
means of accomplishing this objective. If we adopt a Bayesian viewpoint,
then we want the filter to propagate the *conditional probability density* of

$$f_{x(i)|z(1),z(2),\dots,z(i)}(x \mid z_1, z_2, \dots, z_i)$$
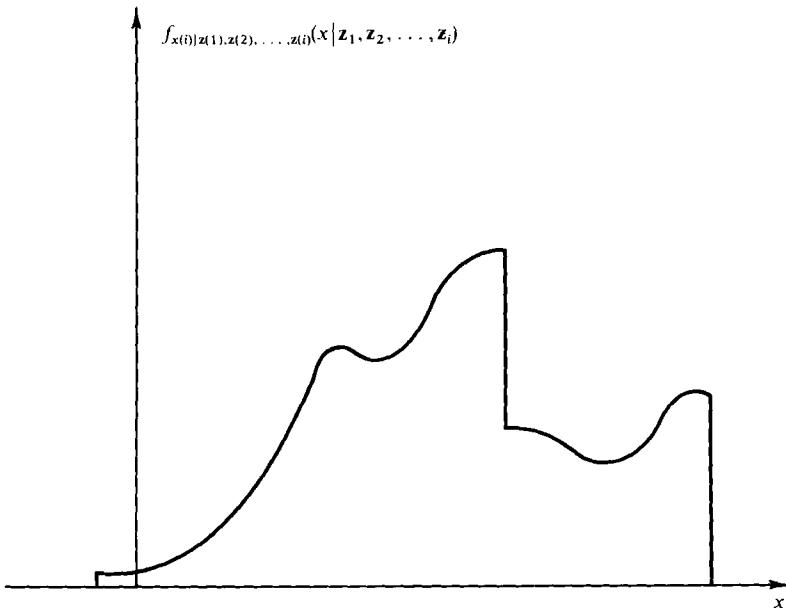
FIG. 1.2   Conditional probability density.

the desired quantities, conditioned on knowledge of the actual data coming from the measuring devices. To understand this concept, consider Fig. 1.2, a portrayal of a conditional probability density of the value of a scalar quantity x at time instant i $(x(i))$, conditioned on knowledge that the vector measurement $z(1)$ at time instant 1 took on the value z_1 $(z(1) = z_1)$ and similarly for instants 2 through i, plotted as a function of possible $x(i)$ values. This is denoted as $f_{x(i)|z(1),z(2),\dots,z(i)}(x \mid z_1, z_2, \dots, z_i)$. For example, let $x(i)$ be the one-dimensional position of a vehicle at time instant i, and let $z(j)$ be a two-dimensional vector describing the measurements of position at time j by two separate radars. Such a conditional probability density contains all the available information about $x(i)$: it indicates, for the given value of all measurements taken up through time instant i, what the probability would be of $x(i)$ assuming any particular value or range of values.

It is termed a "conditional" probability density because its shape and location on the x axis is dependent upon the values of the measurements taken. Its shape conveys the amount of certainty you have in the knowledge of the value of x. If the density plot is a narrow peak, then most of the probability "weight" is concentrated in a narrow band of x values. On the other hand, if the plot has a gradual shape, the probability "weight" is spread over a wider range of x, indicating that you are less sure of its value.

Once such a conditional probability density function is propagated, the "optimal" estimate can be defined. Possible choices would include

(1) the *mean*—the "center of probability mass" estimate;
(2) the *mode*—the value of x that has the highest probability, locating the peak of the density; and
(3) the *median*—the value of x such that half of the probability weight lies to the left and half to the right of it.

A Kalman filter performs this conditional probability density propagation for problems in which the system can be described through a *linear* model and in which system and measurement noises are *white* and *Gaussian* (to be explained shortly). Under these conditions, the mean, mode, median, and virtually any reasonable choice for an "optimal" estimate all coincide, so there is in fact a unique "best" estimate of the value of x. Under these three restrictions, the Kalman filter can be shown to be the best filter of any conceivable form. Some of the restrictions can be relaxed, yielding a qualified optimal filter. For instance, if the Gaussian assumption is removed, the Kalman filter can be shown to be the best (minimum error variance) filter out of the class of linear unbiased filters. However, these three assumptions can be justified for many potential applications, as seen in the following section.

1.4 BASIC ASSUMPTIONS

At this point it is useful to look at the three basic assumptions in the Kalman filter formulation. On first inspection, they may appear to be overly restrictive and unrealistic, To allay any misgivings of this sort, this section will briefly discuss the physical implications of these assumptions.

A linear system model is justifiable for a number of reasons. Often such a model is adequate for the purpose at hand, and when nonlinearities do exist, the typical engineering approach is to linearize about some nominal point or trajectory, achieving a perturbation model or error model. Linear systems are desirable in that they are more easily manipulated with engineering tools, and linear system (or differential equation) theory is much more complete and practical than nonlinear. The fact is that there are means of extending the Kalman filter concept to some nonlinear applications or developing nonlinear filters directly, but these are considered only if linear models prove inadequate.

"Whiteness" implies that the noise value is not correlated in time. Stated more simply, if you know what the value of the noise is now, this knowledge does you no good in predicting what its value will be at any other time. Whiteness also implies that the noise has equal power at all frequencies. Since this results in a noise with infinite power, a white noise obviously cannot really exist. One might then ask, why even consider such a concept if it does not
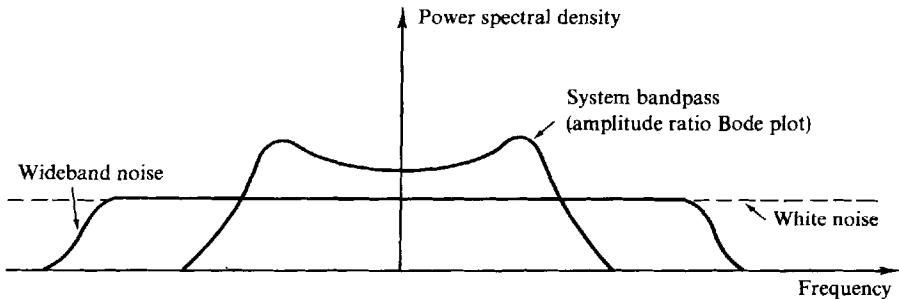
FIG. 1.3   Power spectral density bandwidths.

exist in real life? The answer is twofold. First, any physical system of interest has a certain frequency "bandpass"—a frequency range of inputs to which it can respond. Above this range, the input either has no effect, or the system so severely attentuates the effect that it essentially does not exist. In Fig. 1.3, a typical system bandpass curve is drawn on a plot of "power spectral density" (interpreted as the amount of power content at a certain frequency) versus frequency. Typically a system will be driven by wideband noise—one having power at frequencies above the system bandpass, and essentially constant power at all frequencies within the system bandpass—as shown in the figure. On this same plot, a white noise would merely extend this constant power level out across all frequencies. Now, within the bandpass of the system of interest, the fictitious white noise looks identical to the real wideband noise. So what has been gained? That is the second part of the answer to why a white noise model is used. It turns out that the mathematics involved in the filter are vastly simplified (in fact, made tractable) by replacing the real wideband noise with a white noise which, from the system's "point of view," is identical. Therefore, the white noise model is used.

One might argue that there are cases in which the noise power level is not constant over all frequencies within the system bandpass, or in which the noise is in fact time correlated. For such instances, a white noise put through a small linear system can duplicate virtually any form of time-correlated noise. This small system, called a "shaping filter," is then added to the original system, to achieve an overall linear system driven by white noise once again.

Whereas whiteness pertains to time or frequency relationships of a noise, Gaussianness has to do with its amplitude. Thus, at any single point in time, the probability density of a Gaussian noise amplitude takes on the shape of a normal bell-shaped curve. This assumption can be justified physically by the fact that a system or measurement noise is typically caused by a number of small sources. It can be shown mathematically that when a number of independent random variables are added together, the summed effect can be described very closely by a Gaussian probability density, regardless of the shape of the individual densities.

There is also a practical justification for using Gaussian densities. Similar to whiteness, it makes the mathematics tractable. But more than that, typically an engineer will know, at best, the first and second order statistics (mean and variance or standard deviation) of a noise process. In the absence of any higher order statistics, there is no better form to assume than the Gaussian density. The first and second order statistics completely determine a Gaussian density, unlike most densities which require an endless number of orders of statistics to specify their shape entirely. Thus, the Kalman filter, which propagates the first and second order statistics, includes *all* information contained in the conditional probability density, rather than only some of it, as would be the case with a different form of density.

The particular assumptions that are made are dictated by the objectives of, and the underlying motivation for, the model being developed. If our objective were merely to build good descriptive models, we would not confine our attention to linear system models driven by white Gaussian noise. Rather, we would seek the model, of whatever form, that best fits the data generated by the "real world." It is our desire to build estimators and controllers based upon our system models that drives us to these assumptions: other assumptions generally do not yield tractable estimation or control problem formulations. Fortunately, the class of models that yields tractable mathematics also provides adequate representations for many applications of interest. Later, the model structure will be extended somewhat to enlarge the range of applicability, but the requirement of model usefulness in subsequent estimator or controller design will again be a dominant influence on the manner in which the extensions are made.

1.5 A SIMPLE EXAMPLE

To see how a Kalman filter works, a simple example will now be developed. Any example of a single measuring device providing data on a single variable would suffice, but the determination of a position is chosen because the probability of one's exact location is a familiar concept that easily allows dynamics to be incorporated into the problem.

Suppose that you are lost at sea during the night and have no idea at all of your location. So you take a star sighting to establish your position (for the sake of simplicity, consider a one-dimensional location). At some time t_1 you determine your location to be z_1. However, because of inherent measuring device inaccuracies, human error, and the like, the result of your measurement is somewhat uncertain. Say you decide that the precision is such that the standard deviation (one-sigma value) involved is σ_{z_1} (or equivalently, the variance, or second order statistic, is $\sigma_{z_1}^2$). Thus, you can establish the conditional probability of $x(t_1)$, your position at time t_1, conditioned on the observed value of the measurement being z_1, as depicted in Fig. 1.4. This is a plot of $f_{x(t_1)|z(t_1)}(x|z_1)$ as a function of the location x: it tells you the probability of
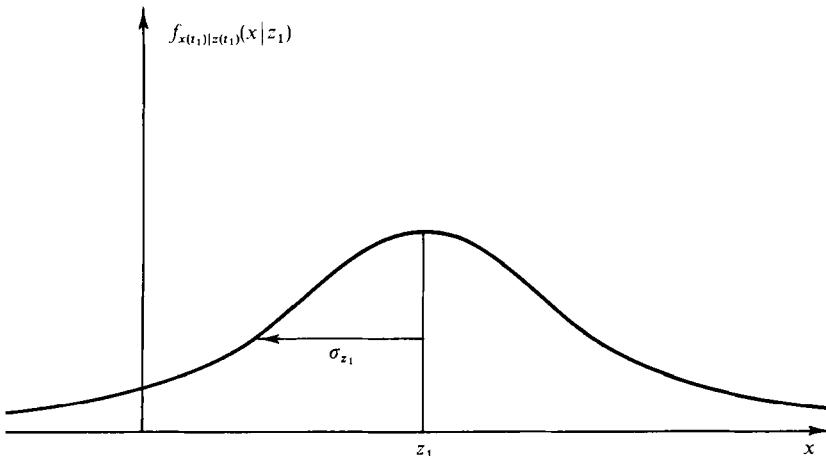
FIG. 1.4 Conditional density of position based on measured value z_1.

being in any one location, based upon the measurement you took. Note that σ_{z_1} is a direct measure of the uncertainty: the larger σ_{z_1} is, the broader the probability peak is, spreading the probability "weight" over a larger range of x values. For a Gaussian density, 68.3% of the probability "weight" is contained within the band σ units to each side of the mean, the shaded portion in Fig. 1.4.

Based on this conditional probability density, the best estimate of your position is

$$\hat{x}(t_1) = z_1 \qquad (1\text{-}1)$$
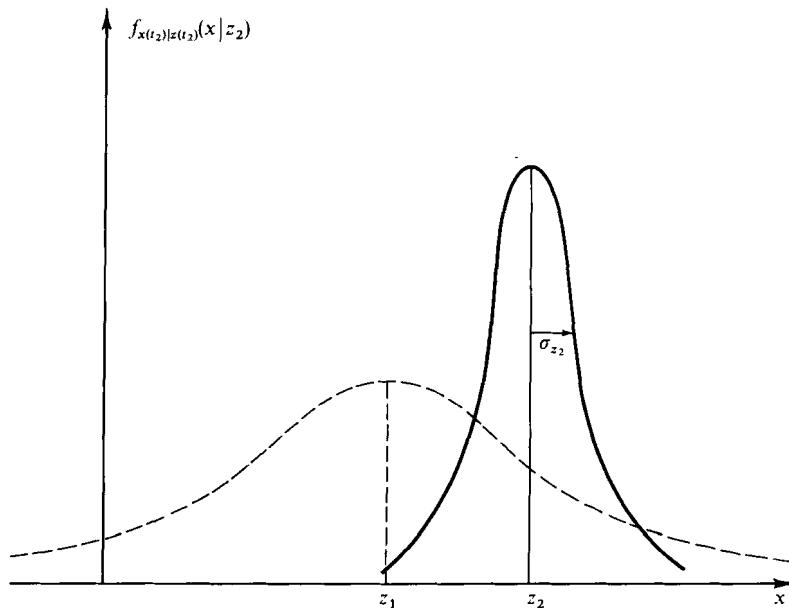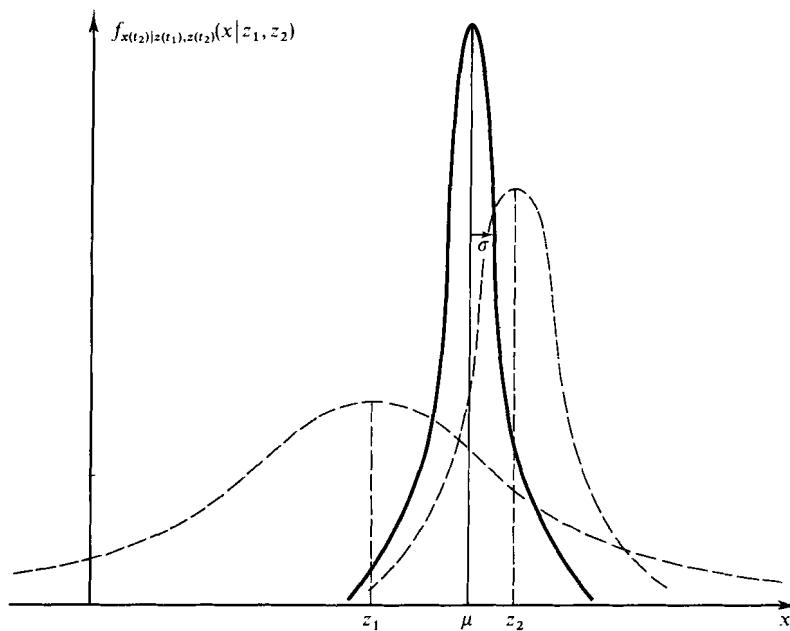
and the variance of the error in the estimate is

$$\sigma_x{}^2(t_1) = \sigma_{z_1}^2 \qquad (1\text{-}2)$$

Note that $\hat{x}$ is both the mode (peak) and the median (value with $\frac{1}{2}$ of the probability weight to each side), as well as the mean (center of mass).

Now say a trained navigator friend takes an independent fix right after you do, at time $t_2 \cong t_1$ (so that the true position has not changed at all), and obtains a measurement z_2 with a variance $\sigma_{z_2}^2$. Because he has a higher skill, assume the variance in his measurement to be somewhat smaller than in yours. Figure 1.5 presents the conditional density of your position at time t_2, based only on the measured value z_2. Note the narrower peak due to smaller variance, indicating that you are rather certain of your position based on his measurement.

At this point, you have two measurements available for estimating your position. The question is, how do you combine these data? It will be shown subsequently that, based on the assumptions made, the conditional density of

FIG. 1.5 Conditional density of position based on measurement z_2 alone.

FIG. 1.6 Conditional density of position based on data z_1 and z_2.

your position at time $t_2 \cong t_1$, $x(t_2)$, given *both* z_1 *and* z_2, is a Gaussian density with mean μ and variance σ^2 as indicated in Fig. 1.6, with

$$\mu = [\sigma_{z_2}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_1 + [\sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_2 \qquad (1\text{-}3)$$

$$1/\sigma^2 = (1/\sigma_{z_1}^2) + (1/\sigma_{z_2}^2) \qquad (1\text{-}4)$$

Note that, from (1-4), σ is less than either σ_{z_1} or σ_{z_2}, which is to say that the uncertainty in your estimate of position has been decreased by combining the two pieces of information.

Given this density, the best estimate is

$$\hat{x}(t_2) = \mu \qquad (1\text{-}5)$$

with an associated error variance σ^2. It is the mode and the mean (or, since it is the mean of a conditional density, it is also termed the conditional mean). Furthermore, it is also the maximum likelihood estimate, the weighted least squares estimate, and the linear estimate whose variance is less than that of any other linear unbiased estimate. In other words, it is the "best" you can do according to just about any reasonable criterion.

After some study, the form of μ given in Eq. (1-3) makes good sense. If σ_{z_1} were equal to σ_{z_2}, which is to say you think the measurements are of equal precision, the equation says the optimal estimate of position is simply the average of the two measurements, as would be expected. On the other hand, if σ_{z_1} were larger than σ_{z_2}, which is to say that the uncertainty involved in the measurement z_1 is greater than that of z_2, then the equation dictates "weighting" z_2 more heavily than z_1. Finally, the variance of the estimate is less than σ_{z_1} even if σ_{z_2} is very large: even poor quality data provide some information, and should thus increase the precision of the filter output.

The equation for $\hat{x}(t_2)$ can be rewritten as

$$\begin{aligned}\hat{x}(t_2) &= [\sigma_{z_2}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_1 + [\sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_2 \\ &= z_1 + [\sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)][z_2 - z_1] \end{aligned} \qquad (1\text{-}6)$$

or, in final form that is actually used in Kalman filter implementations [noting that $\hat{x}(t_1) = z_1$],

$$\hat{x}(t_2) = \hat{x}(t_1) + K(t_2)[z_2 - \hat{x}(t_1)] \qquad (1\text{-}7)$$

where

$$K(t_2) = \sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2) \qquad (1\text{-}8)$$

These equations say that the optimal estimate at time t_2, $\hat{x}(t_2)$, is equal to the best prediction of its value before z_2 is taken, $\hat{x}(t_1)$, plus a correction term of an optimal weighting value times the difference between z_2 and the best prediction of its value before it is actually taken, $\hat{x}(t_1)$. It is worthwhile to understand

this "predictor–corrector" structure of the filter. Based on all previous information, a prediction of the value that the desired variables and measurement will have at the next measurement time is made. Then, when the next measurement is taken, the difference between it and its predicted value is used to "correct" the prediction of the desired variables.

Using the $K(t_2)$ in Eq. (1-8), the variance equation given by Eq. (1-4) can be rewritten as

$$\sigma_x^2(t_2) = \sigma_x^2(t_1) - K(t_2)\sigma_x^2(t_1) \tag{1-9}$$

Note that the values of $\hat{x}(t_2)$ and $\sigma_x^2(t_2)$ embody all of the information in $f_{x(t_2)|z(t_1),\,z(t_2)}(x\,|\,z_1,z_2)$. Stated differently, by propagating these two variables, the conditional density of your position at time t_2, given z_1 and z_2, is completely specified.

Thus we have solved the static estimation problem. Now consider incorporating dynamics into the problem.

Suppose that you travel for some time before taking another measurement. Further assume that the best model you have of your motion is of the simple form

$$dx/dt = u + w \tag{1-10}$$

where u is a nominal velocity and w is a noise term used to represent the uncertainty in your knowledge of the actual velocity due to disturbances, off-nominal conditions, effects not accounted for in the simple first order equation, and the like. The "noise" w will be modeled as a white Gaussian noise with a mean of zero and variance of σ_w^2.

Figure 1.7 shows graphically what happens to the conditional density of position, given z_1 and z_2. At time t_2 it is as previously derived. As time progresses, the density travels along the x axis at the nominal speed u, while simultaneously spreading out about its mean. Thus, the probability density starts at the best estimate, moves according to the nominal model of dynamics,
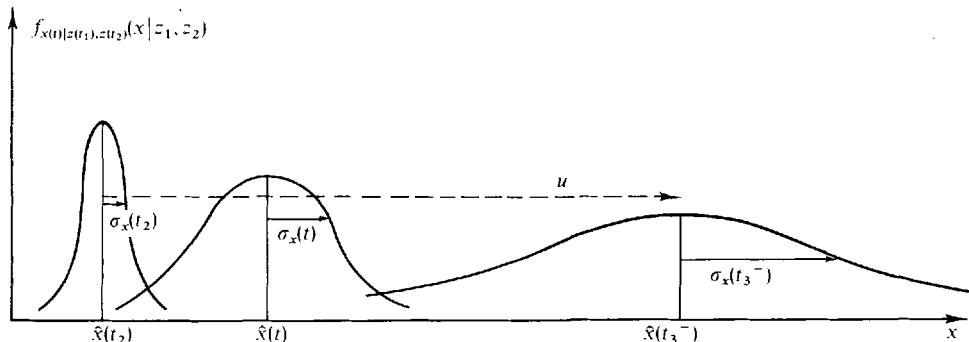
FIG. 1.7   Propagation of conditional probability density.

and spreads out in time because you become less sure of your exact position due to the constant addition of uncertainty over time. At the time t_3^-, just before the measurement is taken at time t_3, the density $f_{x(t_3)|z(t_1),z(t_2)}(x|z_1,z_2)$ is as shown in Fig. 1.7, and can be expressed mathematically as a Gaussian density with mean and variance given by

$$\hat{x}(t_3^-) = \hat{x}(t_2) + u[t_3 - t_2] \tag{1-11}$$

$$\sigma_x^2(t_3^-) = \sigma_x^2(t_2) + \sigma_w^2[t_3 - t_2] \tag{1-12}$$

Thus, $\hat{x}(t_3^-)$ is the optimal prediction of what the x value is at t_3^-, before the measurement is taken at t_3, and $\sigma_x^2(t_3^-)$ is the expected variance in that prediction.

Now a measurement is taken, and its value turns out to be z_3, and its variance is assumed to be $\sigma_{z_3}^2$. As before, there are now two Gaussian densities available that contain information about position, one encompassing all the information available before the measurement, and the other being the information provided by the measurement itself. By the same process as before, the density with mean $\hat{x}(t_3^-)$ and variance $\sigma_x^2(t_3^-)$ is combined with the density with mean z_3 and variance $\sigma_{z_3}^2$, to yield a Gaussian density with mean

$$\hat{x}(t_3) = \hat{x}(t_3^-) + K(t_3)[z_3 - \hat{x}(t_3^-)] \tag{1-13}$$

and variance

$$\sigma_x^2(t_3) = \sigma_x^2(t_3^-) - K(t_3)\sigma_x^2(t_3^-) \tag{1-14}$$

where the gain $K(t_3)$ is given by

$$K(t_3) = \sigma_x^2(t_3^-)/[\sigma_x^2(t_3^-) + \sigma_{z_3}^2] \tag{1-15}$$

The optimal estimate, $\hat{x}(t_3)$, satisfies the same form of equation as seen previously in (1-7). The best prediction of its value before z_3 is taken is corrected by an optimal weighting value times the difference between z_3 and the prediction of its value. Similarly, the variance and gain equations are of the same form as (1-8) and (1-9).

Observe the form of the equation for $K(t_3)$. If $\sigma_{z_3}^2$, the measurement noise variance, is large, then $K(t_3)$ is small; this simply says that you would tend to put little confidence in a very noisy measurement and so would weight it lightly. In the limit as $\sigma_{z_3}^2 \to \infty$, $K(t_3)$ becomes zero, and $\hat{x}(t_3)$ equals $\hat{x}(t_3^-)$: an infinitely noisy measurement is totally ignored. If the dynamic system noise variance σ_w^2 is large, then $\sigma_x^2(t_3^-)$ will be large [see Eq. (1-12)] and so will $K(t_3)$; in this case, you are not very certain of the output of the system model within the filter structure and therefore would weight the measurement heavily. Note that in the limit as $\sigma_w^2 \to \infty$, $\sigma_x^2(t_3^-) \to \infty$ and $K(t_3) \to 1$, so Eq. (1-13) yields

$$\hat{x}(t_3) = \hat{x}(t_3^-) + 1 \cdot [z_3 - \hat{x}(t_3^-)] = z_3 \tag{1-16}$$

Thus in the limit of absolutely no confidence in the system model output, the optimal policy is to ignore the output and use the new measurement as the optimal estimate. Finally, if $\sigma_x^2(t_3^-)$ should ever become zero, then so does $K(t_3)$; this is sensible since if $\sigma_x^2(t_3^-) = 0$, you are absolutely sure of your estimate before z_3 becomes available and therefore can disregard the measurement.

Although we have not as yet derived these results mathematically, we have been able to demonstrate the reasonableness of the filter structure.

1.6 A PREVIEW

Extending Eqs. (1-11) and (1-12) to the vector case and allowing time varying parameters in the system and noise descriptions yields the general Kalman filter algorithm for propagating the conditional density and optimal estimate from one measurement sample time to the next. Similarly, the Kalman filter update at a measurement time is just the extension of Eqs. (1-13)–(1-15). Further logical extensions would include estimation with data beyond the time when variables are to be estimated, estimation with nonlinear system models rather than linear, control of systems described through stochastic models, and both estimation and control when the noise and system parameters are not known with absolute certainty. The sequel provides a thorough investigation of those topics, developing both the theoretical mathematical aspects and practical engineering insights necessary to resolve the problem formulations and solutions fully.

GENERAL REFERENCES

The following references have influenced the development of both this introductory chapter and the entirety of this text.

1. Aoki, M., *Optimization of Stochastic Systems—Topics in Discrete-Time Systems*. Academic Press, New York, 1967.
2. Åström, K. J., *Introduction to Stochastic Control Theory*. Academic Press, New York, 1970.
3. Bryson, A. E. Jr., and Ho, Y., *Applied Optimal Control*. Blaisdell, Waltham, Massachusetts, 1969.
4. Bucy, R. S., and Joseph, P. D., *Filtering for Stochastic Processes with Applications to Guidance*. Wiley, New York, 1968.
5. Deutsch, R., *Estimation Theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1965.
6. Deyst, J. J., "Estimation and Control of Stochastic Processes," unpublished course notes. M.I.T. Dept. of Aeronautics and Astronautics, Cambridge, Massachusetts, 1970.
7. Gelb, A. (ed.), *Applied Optimal Estimation*. M.I.T. Press, Cambridge, Massachusetts, 1974.
8. Jazwinski, A. H., *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.
9. Kwakernaak, H., and Sivan, R., *Linear Optimal Control Systems*. Wiley, New York, 1972.
10. Lee, R. C. K., *Optimal Estimation, Identification and Control*. M.I.T. Press, Cambridge, Massachusetts, 1964.
11. Liebelt, P. B., *An Introduction to Optimal Estimation*. Addison-Wesley, Reading, Massachusetts, 1967.
12. Maybeck, P. S., "The Kalman Filter—An Introduction for Potential Users," TM-72-3. Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, June 1972.

13.   Maybeck, P. S., "Applied Optimal Estimation- -Kalman Filter Design and Implementation," notes for a continuing education course offered by the Air Force Institute of Technology, Wright-Patterson AFB, Ohio, semiannually since December 1974.

14.   Meditch, J. S., *Stochastic Optimal Linear Estimation and Control*. McGraw-Hill, New York, 1969.

15.   McGarty, T. P., *Stochastic Systems and State Estimation*. Wiley, New York, 1974.

16.   Sage, A. P., and Melsa, J. L., *Estimation Theory with Application to Communications and Control*. McGraw-Hill, New York, 1971.

17.   Schweppe, F. C., *Uncertain Dynamic Systems*. Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

18.   Van Trees, H. L., *Detection, Estimation and Modulation Theory*, Vol. 1. Wiley, New York, 1968.

APPENDIX AND PROBLEMS

Matrix Analysis

This appendix and its associated problems present certain results from elementary matrix analysis, as well as notation conventions, that will be of use throughout the text. If the reader desires more than this brief review, the list of references [1–11] at the end provides a partial list of good sources.

A.1   Matrices

An *n-by-m matrix* is a rectangular array of scalars consisting of n rows and m columns, denoted by a boldfaced capitalized letter, as

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{bmatrix}$$

Thus, A_{ij} is the scalar element in the ith row and jth column of $\mathbf{A}$, and unless specified otherwise, will be assumed herein to be a real number (or a real-valued scalar function).

If all of the elements A_{ij} are zeros, $\mathbf{A}$ is called a *zero matrix* or *null matrix*, denoted as $\mathbf{0}$.

If all of the elements of an *n-by-n* (*square*) matrix are zeros except for those along the principal diagonal, as

$$\mathbf{A} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix}$$

the $\mathbf{A}$ is called *diagonal*. Furthermore, if $A_{ii} = 1$ for all i, the matrix is called the *identity matrix* and is denoted by $\mathbf{I}$.

A square matrix is *symmetric* if $A_{ij} = A_{ji}$ for all values of i and j from 1 to n. Thus, a diagonal matrix is always symmetric. Show that there are at most $\frac{1}{2}n(n + 1)$ nonredundant elements in an n-by-n symmetric matrix.

A *lower triangular* matrix is a square matrix, all of whose elements above the principal diagonal are zero, as

$$
\mathbf{A} = \begin{bmatrix}
A_{11} & 0 & \cdots & 0 \\
A_{21} & A_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
A_{n1} & A_{n2} & \cdots & A_{nn}
\end{bmatrix}
$$

Similarly, an *upper triangular* matrix is a square matrix with all zeros below the principal diagonal.

A matrix composed of a single column, i.e., an n-by-1 matrix, is called an n-dimensional *vector* or *n-vector* and will be denoted by a boldfaced lower case letter, as

$$
\mathbf{x} = \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix}
$$

Thus, x_i is the ith scalar element, or "component," of the n-vector $\mathbf{x}$. (The directed line segment from the origin to a point in Euclidean n-dimensional space can be represented, relative to a chosen basis or reference coordinate directions, by $\mathbf{x}$, and then x_i is the component along the ith basis vector or reference direction.) Properties of general nonsquare matrices (as described in Sections A.2, A.3, and A.10 to follow) are true specifically for vectors.

A matrix can be subdivided not only into its scalar elements, but also into arrays of elements called matrix *partitions*, such as

$$
\mathbf{A} = \left[\begin{array}{c|c|c|c}
\mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1i} \\
\vdots & \vdots & & \vdots \\
\hline
\mathbf{A}_{k1} & \mathbf{A}_{k2} & \cdots & \mathbf{A}_{ki}
\end{array}\right]
\begin{array}{l} \} n_1 \text{ rows} \\ \\ \} n_k \text{ rows} \end{array}
\qquad
\mathbf{x} = \begin{bmatrix}
\mathbf{x}_1 \\
\mathbf{x}_2 \\
\vdots \\
\mathbf{x}_k
\end{bmatrix}
\begin{array}{l} \} n_1 \text{ components} \\ \} n_2 \text{ components} \\ \\ \} n_k \text{ components} \end{array}
$$

$\underbrace{}_{\substack{m_1 \\ \text{columns}}} \quad \underbrace{}_{\substack{m_2 \\ \text{columns}}} \qquad \underbrace{}_{\substack{m_i \\ \text{columns}}}$

A square matrix $\mathbf{A}$ is termed *block diagonal* if it can be subdivided into partitions such that $\mathbf{A}_{ij} = 0$ for all partitions for which $i \neq j$, and such that all partitions $\mathbf{A}_{ii}$ are square.

A.2   Equality, Addition, and Multiplication

Two n-by-m matrices $\mathbf{A}$ and $\mathbf{B}$ are *equal* if and only if $A_{ij} = B_{ij}$ for all i and j.

If $\mathbf{A}$ and $\mathbf{B}$ are both n-by-m matrices, their *sum* can be defined as $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where $\mathbf{C}$ is an n-by-m matrix whose elements satisfy $C_{ij} = A_{ij} + B_{ij}$ for all i and j.

Their *difference* would be defined similarly. Show that

(a) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
(b) $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$.
(c) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$.

The product of an n-by-m matrix $\mathbf{A}$ by a scalar b is the n-by-m matrix $\mathbf{C} = b\mathbf{A} = \mathbf{A}b$ composed of elements $C_{ij} = bA_{ij}$.

If $\mathbf{A}$ is n-by-m and $\mathbf{B}$ is m-by-r, then the *product* $\mathbf{C} = \mathbf{AB}$ can be defined as an n-by-r matrix with elements $C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}$. This product can be defined only if the number of columns of $\mathbf{A}$ equals the number of rows of $\mathbf{B}$: only if $\mathbf{A}$ and $\mathbf{B}$ are "conformable" for the product $\mathbf{AB}$. Thus, the ordering in the product is important, and $\mathbf{AB}$ can be described as "premultiplying" $\mathbf{B}$ by $\mathbf{A}$ or "post-multiplying" $\mathbf{A}$ by $\mathbf{B}$. Show that for general conformable matrices

(d) $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$.
(e) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$.
(f) $\mathbf{0A} = \mathbf{A0} = \mathbf{0}$.
(g) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.
(h)   in general, $\mathbf{AB} \neq \mathbf{BA}$, even for $\mathbf{A}$ and $\mathbf{B}$ both square.
(i) $\mathbf{AB} = \mathbf{0}$ in general does not imply that $\mathbf{A}$ or $\mathbf{B}$ is $\mathbf{0}$.

A product of particular importance is that of an n-by-m matrix $\mathbf{A}$ with an m-vector $\mathbf{x}$ to yield an n-vector $\mathbf{y} = \mathbf{Ax}$, with components $y_i = \sum_{j=1}^{m} A_{ij}x_j$. Such a matrix multiplication can be used to represent a linear transformation of a vector. More general functions, not expressible through matrix multiplications, can be written as

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \leftrightarrow \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_m) \\ \vdots \\ f_n(x_1, x_2, \ldots, x_m) \end{bmatrix}$$

Matrix operations upon partitioned matrices obey the same rules of equality, addition, and multiplication, provided that the matrix partitions are conformable. For instance, show that

(j)

$$\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right] + \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \hline \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{array}\right]$$

(k)

$$\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array}\right]$$

A.3 Transposition

The *transpose* of an n-by-m matrix $\mathbf{A}$ is the m-by-n matrix denoted as $\mathbf{A}^T$ that satisfies $A_{ij}^T = A_{ji}$ for all i and j. Thus, transposition can be interpreted as interchanging the roles of rows and columns of a matrix. For example, if $\mathbf{x}$ is an n-vector, $\mathbf{x}^T$ is a 1-by-n matrix, or "row vector:" Show that

   (a) $(\mathbf{A}^T)^T = \mathbf{A}$.
   (b) $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$.
   (c) $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.
   (d)   if $\mathbf{A}$ is a symmetric matrix, $\mathbf{A}^T = \mathbf{A}$.
   (e)   if $\mathbf{x}$ and $\mathbf{y}$ are n-vectors, $\mathbf{x}^T\mathbf{y}$ is a scalar and $\mathbf{xy}^T$ is a square n-by-n matrix; $\mathbf{xx}^T$ is symmetric as well.
   (f)   if $\mathbf{A}$ is a symmetric n-by-n matrix and $\mathbf{B}$ is a general m-by-n matrix, then $\mathbf{C} = \mathbf{BAB}^T$ is a symmetric m-by-m matrix.
   (g)   if $\mathbf{A}$ and $\mathbf{B}$ are both symmetric n-by-n matrices, $(\mathbf{A} + \mathbf{B})$ is also symmetric but $(\mathbf{AB})$ generally is not.
   (h)

$$\begin{bmatrix} \mathbf{A} & \vdots & \mathbf{B} \\ \cdots & + & \cdots \\ \mathbf{C} & \vdots & \mathbf{D} \end{bmatrix}^T = \begin{bmatrix} \mathbf{A}^T & \vdots & \mathbf{C}^T \\ \cdots & + & \cdots \\ \mathbf{B}^T & \vdots & \mathbf{D}^T \end{bmatrix}$$

A.4 Matrix Inversion, Singularity, and Determinants

Given a square matrix $\mathbf{A}$, if there exists a matrix such that both premultiplying and postmultiplying it by $\mathbf{A}$ yields the identity, then this matrix is called the *inverse* of $\mathbf{A}$, and is denoted by $\mathbf{A}^{-1}$: $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. A square matrix that does not possess such an inverse is said to be *singular*. If $\mathbf{A}$ has an inverse, the inverse is unique, and $\mathbf{A}$ is termed *nonsingular*. Show that

   (a)   if $\mathbf{A}$ is nonsingular, then so is $\mathbf{A}^{-1}$, and $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
   (b) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if all indicated inverses exist.
   (c) $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$.
   (d)   if a transformation of variables is represented by $\mathbf{x}^* = \mathbf{Ax}$ and if $\mathbf{A}^{-1}$ exists, then $\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}^*$.

The *determinant* of a square n-by-n matrix $\mathbf{A}$ is a scalar-valued function of the matrix elements, denoted by $|\mathbf{A}|$, the evaluation of which can be performed recursively through $|\mathbf{A}| = \sum_{j=1}^{n} A_{ij}C_{ij}$ for any fixed $i = 1, 2, \ldots,$ or n; C_{ij} is the "cofactor" of A_{ij}, defined through $C_{ij} = (-1)^{i+j}M_{ij}$, and M_{ij} is the "minor" of A_{ij}, defined as the determinant of the $(n-1)$-by-$(n-1)$ matrix formed by deleting the ith row and jth column of the n-by-n $\mathbf{A}$. (Note that iterative application of these relationships ends with the evaluation of determinants of 1-by-1 matrices or scalars as the scalar values themselves.) Show that

   (e)   if $\mathbf{A}$ is 2-by-2, then $|\mathbf{A}| = A_{11}A_{22} - A_{12}A_{21}$.

(f)   if $\mathbf{A}$ is 3-by-3, then

$$|\mathbf{A}| = A_{11}A_{22}A_{33} + A_{12}A_{23}A_{31} + A_{13}A_{32}A_{21}$$
$$- A_{11}A_{32}A_{23} - A_{12}A_{21}A_{33} - A_{13}A_{22}A_{31}$$

(g) $|\mathbf{A}^{\mathrm{T}}| = |\mathbf{A}|$.

(h)   if all the elements of any row or column of $\mathbf{A}$ are zero, $|\mathbf{A}| = 0$.

(i)   if any row (column) of $\mathbf{A}$ is a multiple of any other row (column), then $|\mathbf{A}| = 0$.

(j)   if a scalar multiple of any row (column) is added to any other row (column) of $\mathbf{A}$, the value of the determinant is unchanged.

(k)   if $\mathbf{A}$ and $\mathbf{B}$ are n-by-n, $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.

(l)   if $\mathbf{A}$ is diagonal, then $|\mathbf{A}|$ equals the product of its diagonal elements: $|\mathbf{A}| = \prod_{i=1}^{n} A_{ii}$.

(m)   if the n-by-n $\mathbf{A}$ is nonsingular, then $|\mathbf{A}| \neq 0$ and $\mathbf{A}^{-1}$ can be evaluated as $\mathbf{A}^{-1} = [\mathrm{adj}\,\mathbf{A}]/|\mathbf{A}|$, where $[\mathrm{adj}\,\mathbf{A}]$ is the adjoint of $\mathbf{A}$, defined as the n-by-n matrix whose ij element (i.e., in the ith row and jth column) is the cofactor C_{ji}.

(n) $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$ if $|\mathbf{A}| \neq 0$.

(o) $|\mathbf{A}| = 0$ if and only if $\mathbf{A}$ is singular.

(p)

$$\left| \begin{bmatrix} \mathbf{A} & \vdots & \mathbf{B} \\ \hline \mathbf{0} & \vdots & \mathbf{C} \end{bmatrix} \right| = |\mathbf{A}|\,|\mathbf{C}|$$

If $\mathbf{A}$ is such that its inverse equals its transpose, $\mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}}$, then $\mathbf{A}$ is termed *orthogonal*. If $\mathbf{A}$ is orthogonal, $\mathbf{AA}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$, and $|\mathbf{A}| = \pm 1$.

A.5   Linear Independence and Rank

A set of k n-vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is said to be *linearly dependent* if there exists a set of k constants $c_1, c_2, \ldots, c_k$ (at least one of which is not zero) such that $\sum_{i=1}^{k} c_i \mathbf{x}_i = \mathbf{0}$. If no such set of constants exists, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are said to be *linearly independent*.

The *rank* of an n-by-m matrix is the order of the largest square nonsingular matrix that can be formed by deleting rows and columns. Show that

(a)   if the m-by-n $\mathbf{A}$ is partitioned into column vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ and $\mathbf{x}$ is an n-vector, then $\mathbf{Ax} = \sum_{i=1}^{n} \mathbf{a}_i x_i$.

(b)   the rank of $\mathbf{A}$ equals the number of linearly independent rows or columns of $\mathbf{A}$, whichever is smaller.

(c)   if $\mathbf{A}$ is n-by-n, then it is of rank n (of "full rank") if and only if it is nonsingular.

(d)   the rank of $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ is one.

A.6   Eigenvalues and Eigenvectors

The equation $\mathbf{Ax} = \lambda\mathbf{x}$ for n-by-n $\mathbf{A}$, or $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, possesses a nontrivial solution if an only if $|\mathbf{A} - \lambda\mathbf{I}| = 0$. The nth order polynomial $f(\lambda) = |\mathbf{A} - \lambda\mathbf{I}|$ is called the *characteristic polynomial* of $\mathbf{A}$, and the equation $f(\lambda) = 0$ is called its *characteristic equation*. The n eigenvalues of $\mathbf{A}$ are the (not necessarily distinct) roots of this equation, and the nonzero solutions to $\mathbf{Ax}_i = \lambda_i\mathbf{x}_i$, corresponding to the roots λ_i, are called *eigenvectors*. It can be shown that $|\mathbf{A}|$ equals the product of the eigenvalues of $\mathbf{A}$, and $\sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i$.

Let the eigenvalues of the n-by-n $\mathbf{A}$ be the distinct values $\lambda_1, \lambda_2, \ldots, \lambda_n$, and let the associated eigenvectors be $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. Then, if $\mathbf{E} = [\mathbf{e}_1 | \mathbf{e}_2 | \cdots | \mathbf{e}_n]$, $\mathbf{E}$ is nonsingular, and $\mathbf{E}^{-1}\mathbf{A}\mathbf{E}$ is a diagonal matrix whose ith diagonal element is λ_i, for $i = 1, 2, \ldots, n$. Moreover, if $\mathbf{A}$ is also symmetric, then the eigenvalues are all real and $\mathbf{E}$ is orthogonal.

(a)   Obtain the eigenvalues and eigenvectors for a general 2-by-2 $\mathbf{A}$; generate $\mathbf{E}$ and $\mathbf{E}^{-1}\mathbf{A}\mathbf{E}$.

(b)   Repeat for a general symmetric 2-by-2 $\mathbf{A}$; show that λ_1 and λ_2 must be real, and that $\mathbf{E}$ is orthogonal.

(c)   Show that $|\mathbf{A}| = 0$ if and only if at least one eigenvalue is zero.

A.7   Quadratic Forms and Positive (Semi-) Definiteness

If $\mathbf{A}$ is n-by-n and $\mathbf{x}$ is an n-vector, then the scalar quantity $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is called a *quadratic form*. Show that

(a) $\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}x_i x_j$.

(b)   without loss of generality, $\mathbf{A}$ can always be considered to be symmetric, since if $\mathbf{A}$ is not symmetric, a symmetric matrix $\mathbf{B}$ can always be defined by

$$B_{ij} = \begin{cases} A_{ij} & i = j \\ \frac{1}{2}(A_{ij} + A_{ji}) & i \neq j \end{cases}$$

for i and j equal to $1, 2, \ldots, n$, and then $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{B}\mathbf{x}$.

(c)   if $\mathbf{A}$ is diagonal, $\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} A_{ii}x_i^2$.

If $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, the quadratic form is said to be *positive definite*, as is the matrix $\mathbf{A}$ itself, often written notationally as $\mathbf{A} > 0$. If $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$, the quadratic form and matrix $\mathbf{A}$ are termed *positive semidefinite*, denoted as $\mathbf{A} \geq 0$. Furthermore, the notation $\mathbf{A} > \mathbf{B}$ $(\mathbf{A} \geq \mathbf{B})$ is meant to say that $(\mathbf{A} - \mathbf{B})$ is positive definite (semidefinite). Show that

(d)   if $\mathbf{A}$ is positive definite, it is nonsingular, and its inverse $\mathbf{A}^{-1}$ is also positive definite.

(e)   the symmetric **A** is positive definite if and only if its eigenvalues are all positive; **A** is positive semidefinite if and only if its eigenvalues are all positive or zero.

(f)   if **A** is positive definite and **B** is positive semidefinite, (**A** + **B**) is positive definite.

A.8   Trace

The *trace* of an n-by-n matrix **A**, denoted as tr(**A**), is defined as the sum of the diagonal terms:

$$\text{tr}(\mathbf{A}) \triangleq \sum_{i=1}^{n} A_{ii}$$

Using this basic definition, show that

(a) $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^{\mathsf{T}})$.

(b) $\text{tr}(\mathbf{A}_1 + \mathbf{A}_2) = \text{tr}(\mathbf{A}_1) + \text{tr}(\mathbf{A}_2)$.

(c)   if **B** is n-by-m and **C** is m-by-n, so that **BC** is n-by-n and **CB** is m-by-m, then

$$\text{tr}(\mathbf{BC}) = \text{tr}(\mathbf{CB}) = \text{tr}(\mathbf{B}^{\mathsf{T}}\mathbf{C}^{\mathsf{T}}) = \text{tr}(\mathbf{C}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}})$$

(d)   if **x** and **y** are n-vectors and **A** is n-by-n, then

$$\text{tr}(\mathbf{xy}^{\mathsf{T}}) = \text{tr}(\mathbf{x}^{\mathsf{T}}\mathbf{y}) = \mathbf{x}^{\mathsf{T}}\mathbf{y}$$
$$\text{tr}(\mathbf{Axy}^{\mathsf{T}}) = \text{tr}(\mathbf{y}^{\mathsf{T}}\mathbf{Ax}) = \mathbf{y}^{\mathsf{T}}\mathbf{Ax} = \mathbf{x}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{y}$$

A.9   Similarity

If **A** and **B** are n-by-n and **T** is a nonsingular n-by-n matrix, and $\mathbf{A} = \mathbf{T}^{-1}\mathbf{BT}$, then **A** and **B** are said to be related by a *similarity transformation*, or are simply termed *similar*. Show that

(a)   if $\mathbf{A} = \mathbf{T}^{-1}\mathbf{BT}$, then $\mathbf{B} = \mathbf{TAT}^{-1}$.

(b)   if **A** and **B** are similar, their determinants, eigenvalues, eigenvectors, characteristic polynomials, and traces are equal; also if **A** is positive definite, so is **B** and vice versa.

A.10   Differentiation and Integration

Let **A** be an n-by-m matrix function of a scalar variable t, such as time. Then $d\mathbf{A}/dt \triangleq \dot{\mathbf{A}}(t)$ is defined as the n-by-m matrix with ij element as dA_{ij}/dt for all i and j; $\int \mathbf{A}(\tau)\,d\tau$ is defined similarly as a matrix composed of elements $\int A_{ij}(\tau)\,d\tau$. Derivatives and integrals of vectors are special cases of these definitions. Show

that

(a) $d[\mathbf{A}^\mathsf{T}(t)]/dt = [d\mathbf{A}(t)/dt]^\mathsf{T}$ and similarly for integration.

(b) $d[\mathbf{A}(t)\mathbf{B}(t)]/dt = \dot{\mathbf{A}}(t)\mathbf{B}(t) + \mathbf{A}(t)\dot{\mathbf{B}}(t)$.

Let the scalar s and the n-vector $\mathbf{x}$ be functions of the m-vector $\mathbf{v}$. By convention, the following derivative definitions are made:

$$\frac{\partial s}{\partial \mathbf{v}} = \begin{bmatrix} \dfrac{\partial s}{\partial v_1} & \dfrac{\partial s}{\partial v_2} & \cdots & \dfrac{\partial s}{\partial v_m} \end{bmatrix}$$

$$\frac{\partial \mathbf{x}}{\partial \mathbf{v}} = \begin{bmatrix} \dfrac{\partial \mathbf{x}}{\partial v_1} & \vdots & \dfrac{\partial \mathbf{x}}{\partial v_2} & \vdots & \cdots & \vdots & \dfrac{\partial \mathbf{x}}{\partial v_m} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial x_1}{\partial v_1} & \dfrac{\partial x_1}{\partial v_2} & \cdots & \dfrac{\partial x_1}{\partial v_m} \\[2mm] \dfrac{\partial x_2}{\partial v_1} & \dfrac{\partial x_2}{\partial v_2} & \cdots & \dfrac{\partial x_2}{\partial v_m} \\[2mm] \vdots & \vdots & \ddots & \vdots \\[2mm] \dfrac{\partial x_n}{\partial v_1} & \dfrac{\partial x_n}{\partial v_2} & \cdots & \dfrac{\partial x_n}{\partial v_m} \end{bmatrix}$$

By generating the appropriate forms for scalar components and recombining, show the validity of the following useful forms (for the vectors $\mathbf{x}$ and $\mathbf{y}$ assumed to be functions of $\mathbf{v}$ possibly, and the vector $\mathbf{z}$ and matrices $\mathbf{A}$ and $\mathbf{B}$ assumed constant):

(c) $\partial \mathbf{v}/\partial \mathbf{v} = \mathbf{I}$.

(d) $\partial(\mathbf{Ax})/\partial \mathbf{v} = \mathbf{A}\,\partial \mathbf{x}/\partial \mathbf{v}$, and thus, $\partial(\mathbf{Av})/\partial \mathbf{v} = \mathbf{A}$.

(e) $\partial(\mathbf{x}^\mathsf{T}\mathbf{Ay})/\partial \mathbf{v} = \mathbf{x}^\mathsf{T}\mathbf{A}\,\partial \mathbf{y}/\partial \mathbf{v} + \mathbf{y}^\mathsf{T}\mathbf{A}^\mathsf{T}\,\partial \mathbf{x}/\partial \mathbf{v}$

and so

$$\partial(\mathbf{z}^\mathsf{T}\mathbf{Av})/\partial \mathbf{v} = \mathbf{z}^\mathsf{T}\mathbf{A}, \qquad \partial(\mathbf{v}^\mathsf{T}\mathbf{Az})/\partial \mathbf{v} = \mathbf{z}^\mathsf{T}\mathbf{A}^\mathsf{T}$$

and

$$\partial(\mathbf{v}^\mathsf{T}\mathbf{Av})/\partial \mathbf{v} = \mathbf{v}^\mathsf{T}\mathbf{A} + \mathbf{v}^\mathsf{T}\mathbf{A}^\mathsf{T} = 2\mathbf{v}^\mathsf{T}\mathbf{A} \qquad \text{if} \quad \mathbf{A} = \mathbf{A}^\mathsf{T}$$

and

$$\partial\{(\mathbf{z} - \mathbf{Bv})^\mathsf{T}\mathbf{A}(\mathbf{z} - \mathbf{Bv})\}/\partial \mathbf{v} = -2(\mathbf{z} - \mathbf{Bv})^\mathsf{T}\mathbf{AB} \qquad \text{if} \quad \mathbf{A} = \mathbf{A}^\mathsf{T}.$$

REFERENCES

1. Bellman, R. E., *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1960.
2. DeRusso, P. M., Roy, R. J., and Close, C. M., *State Variables for Engineers*. Wiley, New York, 1965.
3. Edelen, D. G. B., and Kydoniefs, A. D., *An Introduction to Linear Algebra for Science and Engineering* (2nd Ed.). American Elsevier, New York, 1976.

4.   Gantmacher, R. F., *Matrix Theory*, Vol. I, Chelsea, New York, 1959 (translation of 1953 Russian edition).
5.   Hildebrand, F. B., *Methods of Applied Mathematics* (2nd Ed.). Prentice-Hall, Englewood Cliffs, New Jersey, 1965.
6.   Hoffman, K., and Kunze, R., *Linear Algebra*. Prentice-Hall, Englewood Cliffs, New Jersey, 1961.
7.   Nering, E. D., *Linear Algebra and Matrix Theory* (2nd Ed.). Wiley, New York, 1970.
8.   Noble, B., *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969.
9.   Ogata, K., *State Space Analysis of Control Systems*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.
10.   Polak, E., and Wong, E., *Notes for a First Course on Linear Systems*. Van Nostrand-Reinhold, Princeton, New Jersey, 1970.
11.   Zadeh, L. A., and Desoer, C. A., *Linear System Theory*. McGraw-Hill, New York, 1963.

CHAPTER **2**
Deterministic system models

2.1 INTRODUCTION

This chapter reviews the basics of deterministic system modeling, emphasizing time-domain methods. A strong foundation in deterministic models provides invaluable insights into, and motivation for, stochastic models to be developed subsequently. Especially because the ability to generate adequate models for a given application will typically be the critical factor in designing a practical estimation or control algorithm, considerably more detail will be developed herein than might be expected of a review.

Section 2.2 develops continuous-time dynamic models, perhaps the most natural description of most problems of practical interest. Attention progresses from linear, time-invariant, single input–single output systems models through nonlinear state models, exploiting as many analytical tools as practical to gain insights. Solutions to the state differential equations in these models are then discussed in Section 2.3. Because estimators and controllers will eventually be implemented digitally in most cases, discrete-time measurement outputs from continuous-time systems are investigated in Section 2.4. Finally, the properties of controllability and observability are discussed in Section 2.5.

2.2 CONTINUOUS-TIME DYNAMIC MODELS

Our basic models of dynamic systems will be in the form of state space representations [1–19]. These time-domain models will be emphasized because they can address a more general class of problems than frequency domain model formulations, but the useful interrelationships between the two forms will also be exploited for those problems in which both are valid. This section will begin with the simplest model forms and generalize to the most general model of a dynamic system.

Let us first restrict our attention to *linear, time-invariant, single input–single output system models,* those which can be adequately described by means of a linear constant-coefficient ordinary nth order differential equation of the form

$$\frac{d^n z(t)}{dt^n} + a_{n-1} \frac{d^{n-1} z(t)}{dt^{n-1}} + \cdots + a_0 z(t) = c_p \frac{d^p u(t)}{dt^p} + \cdots + c_0 u(t) \qquad (2\text{-}1)$$

where $u(t)$ is the system input at time t and $z(t)$ is the corresponding system output. Because of the linear, time-invariant structure, we can take the Laplace transform of Eq. (2-1), and rearrange (letting initial conditions be zero, but this can be generalized) to generate the system *transfer function* $G(s)$ to describe the output:

$$z(s) = G(s)u(s) \qquad (2\text{-}2)$$

$$G(s) = \frac{c_p s^p + c_{p-1} s^{p-1} + \cdots + c_1 s + c_0}{s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0} \qquad (2\text{-}3)$$

The denominator of $G(s)$ reveals that we have an nth order system model, i.e., the homogeneous differential equation is of order n. The dynamic behavior of the system can be described by the *poles* of $G(s)$—the roots of this denominator.

A corresponding *state space representation* of the same system (for $n > p$) would be a first order vector differential equation with associated output relation:

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{b}u(t) \qquad (2\text{-}4)$$

$$z(t) = \mathbf{h}^{\mathrm{T}}\mathbf{x}(t) \qquad (2\text{-}5)$$

Here $\mathbf{x}$ is an n-dimensional state vector (the n dimensions corresponding to the fact that the system is described by nth order dynamics), the overdot denotes time derivative, $\mathbf{F}$ is a constant n-by-n matrix, $\mathbf{b}$ and $\mathbf{h}$ are constant n-dimensional vectors, and $^{\mathrm{T}}$ denotes transpose ($\mathbf{h}^{\mathrm{T}}$ is thus a 1-by-n matrix).

The state vector is a set of n variables, the values of which are sufficient to describe the system behavior completely. To be more precise, the *state* of a system at any time t is a minimum set of values $x_1(t), \ldots, x_n(t)$, which, along with the input to the system for all time τ, $\tau \geq t$, is sufficient to determine the behavior of the system for all $\tau \geq t$. In order to specify the solution to an nth order differential equation completely, we must prescribe n initial conditions at the initial time t_0 and the forcing function for all $t \geq t_0$ of interest: there are n quantities required to establish the system "state" at t_0. But t_0 can be any time of interest, so we can see that n variables are required to establish the state at any given time.

EXAMPLE 2.1   Consider the one-dimensional motion of a point mass on an ideal spring. If we specify the forcing function and the initial conditions on the position and velocity of the mass, we can completely describe the future behavior of the system through the solution of a second order differential equation. In this case, the state vector is two dimensional.   ∎

It can be shown that the relation between the state space representation given by (2-4) and (2-5) and the transfer function given by (2-3) is

$$G(s) = \mathbf{h}^T[s\mathbf{I} - \mathbf{F}]^{-1}\mathbf{b} \qquad (2\text{-}6)$$

The matrix $[s\mathbf{I} - \mathbf{F}]^{-1}$ is often given the symbol $\mathbf{\Phi}(s)$ and called the resolvent matrix; it is the Laplace transform of the system state transition matrix, to be discussed in the next section. Equation (2-6) reveals the fact that the poles of the transfer function, the defining parameters of the homogeneous system, are equal to the eigenvalues of the matrix $\mathbf{F}$. Given an nth order transfer function model with no pole-zero cancellations, it is possible to generate an n-dimensional state representation that duplicates its input–output characteristics. It is also possible to generate "nonminimal" state representations, of order greater than n, by adding extraneous or redundant variables to the state vector. However, such representations cannot be both observable and controllable, concepts to be discussed in Section 2.4. Figure 2.1 depicts these equivalent system models.
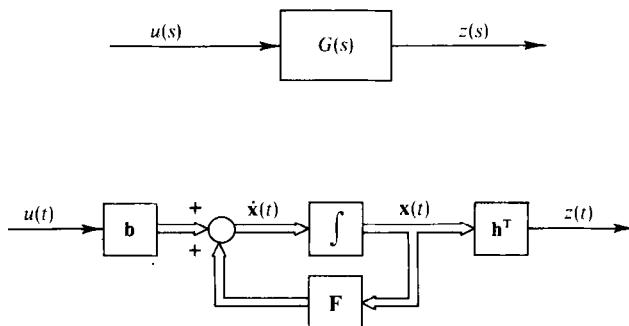
FIG. 2.1   Equivalent system representations. (⇒ denotes vector quantities in this figure.)

However, the state space representation is not unique. A certain set of state variables uniquely determines the system behavior, but there is an infinite number of such sets. There are four major types of state space representations: physical, standard controllable, standard observable, and canonical variables [1–3, 6, 15–17, 19]. Of these, the first two are the most readily generated: in the first, the states are physical quantities and Eqs. (2-4) and (2-5) result from combining the relationships (physical "laws," etc.) among these physical quantities; the second is particularly simple to derive from a previously determined transfer function model. The canonical form decouples the modes of the system dynamics and thereby facilitates both system analysis and numerical solutions to the state differential equation.

The various equivalent state representations can be related through *similarity transformations* (geometrically defining new basis vectors in n-dimensional

state space). Given a system described by (2-4) and (2-5):

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{b}u(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0; \qquad z(t) = \mathbf{h}^T\mathbf{x}(t)$$

we can define a new state vector $\mathbf{x}^*(t)$ through an invertible n-by-n matrix $\mathbf{T}$ as

$$\mathbf{x}(t) = \mathbf{T}\mathbf{x}^*(t) \tag{2-7a}$$

$$\mathbf{x}^*(t) = \mathbf{T}^{-1}\mathbf{x}(t) \tag{2-7b}$$

Substituting (2-7) into (2-4) yields

$$\mathbf{T}\dot{\mathbf{x}}^*(t) = \mathbf{F}\mathbf{T}\mathbf{x}^*(t) + \mathbf{b}u(t)$$

Premultiplying this by $\mathbf{T}^{-1}$, and substituting (2-7a) into (2-5) generates the result as

$$\dot{\mathbf{x}}^*(t) = \mathbf{F}^*\mathbf{x}^*(t) + \mathbf{b}^*u(t); \qquad \mathbf{x}^*(t_0) = \mathbf{x}_0^* \tag{2-8}$$

$$z(t) = \mathbf{h}^{*T}\mathbf{x}^*(t) \tag{2-9}$$

where

$$\mathbf{F}^* = \mathbf{T}^{-1}\mathbf{F}\mathbf{T} \tag{2-10a}$$

$$\mathbf{b}^* = \mathbf{T}^{-1}\mathbf{b} \tag{2-10b}$$

$$\mathbf{h}^{*T} = \mathbf{h}^T\mathbf{T} \tag{2-10c}$$

Under a similarity transformation such as (2-10a), the eigenvalues, determinant, trace, and characteristic polynomial are all invariant. Thus, since the eigenvalues, i.e., system poles, remain unchanged, the transformation has not altered the dynamics of the system representation.

   Physical variables are desirable for applications where feedback is needed, since they are often directly measurable. However, there is no uniformity in the structures of $\mathbf{F}$, $\mathbf{b}$, and $\mathbf{h}$, and little more can be said about this form without more knowledge of the specific dynamic system.

   The *standard controllable form* can be generated directly from a transfer function or differential equation. Given either (2-1) or (2-3), one can write

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0\cdots 0 & & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2-11}$$

$$z(t) = \begin{bmatrix} c_0 & c_1 & \cdots & c_p & 0 & \cdots & 0 \end{bmatrix}\mathbf{x}(t) \tag{2-12}$$
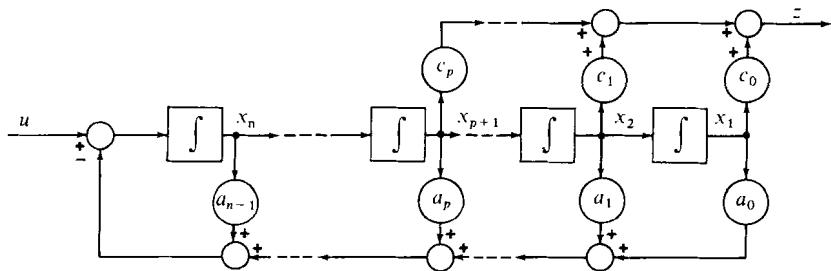
FIG. 2.2   Standard controllable form.

Figure 2.2 presents a block diagram of the standard controllable form state model.

EXAMPLE 2.2   Let a system be described by the transfer function

$$G(s) = \frac{\tau s + 1}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

Here we identify $a_0 = \omega_n^2$, $a_1 = 2\zeta\omega_n$, $c_0 = 1$, $c_1 = \tau$, to yield

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\zeta\omega_n \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

$$z(t) = \begin{bmatrix} 1 & \tau \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

which can be portrayed by the block diagram in Fig. 2.3. Note that the state variables are the outputs of integrators, so that the corresponding inputs directly represent the differential equations. For instance, the output of the left integrator is $x_2(t)$, so the input is

$$\dot{x}_2(t) = -\omega_n^2 x_1(t) - 2\zeta\omega_n x_2(t) + u(t) \qquad \blacksquare$$

The standard controllable form derives its name from the fact that if a non-minimal representation is put into a form of this type, it will be a controllable, but not observable, representation (see Section 2.4).
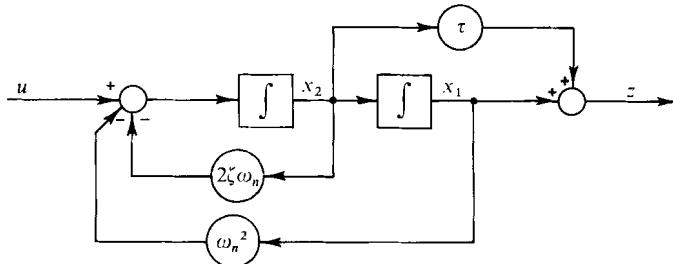
FIG. 2.3   Standard controllable form for Example 2.2.

The *standard observable form* derives its name analogously, and for minimal representations it is described by the same $\mathbf{F}$ matrix as in the standard controllable form, but has different forms for $\mathbf{b}$ and $\mathbf{h}$:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} u(t); \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (2\text{-}13)$$

$$z(t) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \mathbf{x}(t) \qquad (2\text{-}14)$$

Again $\mathbf{F}$ is derived by inspection from either the differential equation or transfer function, and the b_i's are obtained by long division to generate the Laurent series for $G(s)$:

$$G(s) = b_1 s^{-1} + b_2 s^{-2} + \cdots + b_n s^{-n} + \cdots \qquad (2\text{-}15)$$

Figure 2.4 depicts the block diagram for this form.
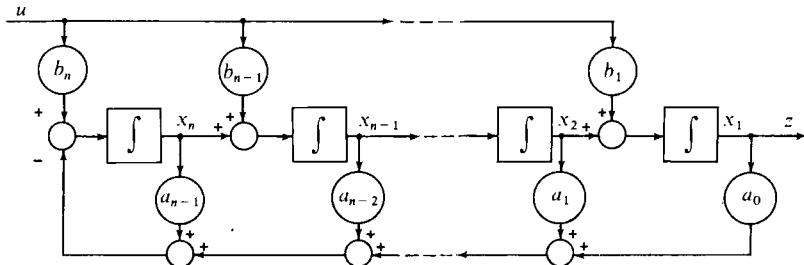
FIG. 2.4   Standard observable form.

EXAMPLE 2.3   Consider the same transfer function as in Example 2.2:

$$G(s) = \frac{\tau s + 1}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

Again $a_0 = \omega_n^2$ and $a_1 = 2\zeta\omega_n$. The b_i values are derived from

$$s^2 + 2\zeta\omega_n s + \omega_n^2 \,\overline{\vert \begin{array}{l} \tau s^{-1} + (1 - 2\zeta\omega_n\tau)s^{-2} + \cdots \\ \tau s + 1 \\ \underline{\tau s + 2\zeta\omega_n\tau + \omega_n^2 \tau s^{-1}} \end{array}}$$

so that $b_1 = \tau$, $b_2 = (1 - 2\zeta\omega_n\tau)$, and the desired result is

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n 2 & -2\zeta\omega_n \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} \tau \\ (1 - 2\zeta\omega_n\tau) \end{bmatrix} u(t)$$

$$z(t) = x_1(t)$$

The associated block diagram is given in Fig. 2.5.   ■
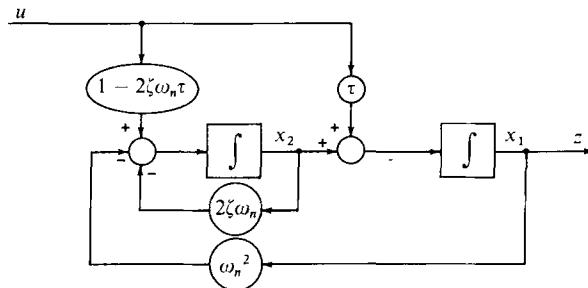
FIG. 2.5   Standard observable form for Example 2.3.

Canonical form provides decoupled system modes; the **F** matrix in this representation is a diagonal matrix whose entries are the eigenvalues of the system, if these eigenvalues are distinct:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} u(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \qquad (2\text{-}16)$$

$$z(t) = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \mathbf{x}(t) \qquad (2\text{-}17)$$

The block diagram is portrayed in Fig. 2.6, from which the separation of system modes is evident.

To obtain this form from a transfer function, the n roots (eigenvalues) of the characteristic polynomial are determined, and $G(s)$ written in terms of the
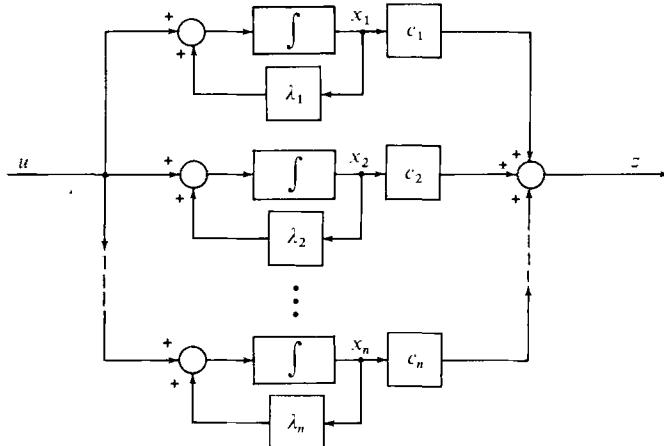
FIG. 2.6   Canonical form state model.

partial fraction expansion

$$G(s) = \frac{r_1}{s - \lambda_1} + \frac{r_2}{s - \lambda_2} + \cdots + \frac{r_n}{s - \lambda_n} \tag{2-18}$$

where λ_i is the ith root (eigenvalue), and r_i is the corresponding residue, given by

$$r_i = (s - \lambda_i)G(s)\Big|_{s = \lambda_i} \tag{2-19}$$

If we let $c_i = r_i$ for $i = 1, 2, \ldots, n$, then the canonical form in (2-16) and (2-17) is completely determined.

EXAMPLE 2.4   Consider the transfer function

$$G(s) = \frac{s + 8}{s^2 + 8s + 12} = \frac{s + 8}{(s + 2)(s + 6)}$$

This can be written as

$$G(s) = \frac{r_1}{s + 2} + \frac{r_2}{s + 6}$$

where

$$r_1 = \frac{s + 8}{s + 6}\Big|_{s = -2} = \frac{3}{2}, \quad r_2 = \frac{s + 8}{s + 2}\Big|_{s = -6} = -\frac{1}{2}$$

yielding a canonical form representation of

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & -6 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t)$$

$$z(t) = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \quad \blacksquare$$

If it is desired to transform any state space representation into canonical variables, first the eigenvalues of the original $\mathbf{F}$ matrix are determined as solutions to

$$|\lambda \mathbf{I} - \mathbf{F}| = 0 \tag{2-20}$$

where $|\cdot|$ denotes determinant. To evaluate the c_i coefficients in (2-17), one can explicitly evaluate the transformation matrix $\mathbf{T}$ [explicit knowledge of $\mathbf{T}$ is also required to transform the initial conditions by $\mathbf{x}^*(t_0) = \mathbf{T}^{-1}\mathbf{x}(t_0)$]. The $\mathbf{F}$, $\mathbf{b}$, and $\mathbf{h}$ in the original representation are known; $\mathbf{F}^*$ for the canonical form is the diagonal matrix of eigenvalues, and $\mathbf{b}^*$ is an n-vector of ones. With such knowledge, the similarity transformation relations (2-10a) and (2-10b) can be written as

$$\mathbf{T}\mathbf{F}^* = \mathbf{F}\mathbf{T} \tag{2-21a}$$

$$\mathbf{T}\mathbf{b}^* = \mathbf{b} \tag{2-21b}$$

and solved simultaneously for **T**. (These are a set of $n^2 + n$ equations, of which n^2 are independent.) Once the transformation matrix **T** is obtained, the desired $\mathbf{h}^{*T}$ is found from

$$\mathbf{h}^{*T} = \mathbf{h}^T\mathbf{T} \tag{2-22}$$

Other means are possible, such as **T** being generated by arraying n eigenvectors in an n-by-n matrix, or by letting **T** be the Vandermonde matrix and not insist **b** be composed of all ones for the case of transforming from standard observable or standard controllable form [5].

EXAMPLE 2.5   Consider the system described in Example 2.4, equivalently modeled by the standard controllable form

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -12 & -8 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t)$$

$$z(t) = \begin{bmatrix} 8 & 1 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

The eigenvalues of **F** are the solutions to

$$|\lambda\mathbf{I} - \mathbf{F}| = \begin{vmatrix} \lambda & -1 \\ 12 & \lambda + 8 \end{vmatrix} = \lambda^2 + 8\lambda + 12 = 0$$

[i.e., the poles of $G(s)$, roots of the characteristic polynomial] from which we obtain $\lambda_1 = -2$, $\lambda_2 = -6$. Then (2-21) becomes:

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}\begin{bmatrix} -2 & 0 \\ 0 & -6 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -12 & -8 \end{bmatrix}\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Solving these yields

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$$

Thus, (2-22) gives $\mathbf{h}^{*T}$ as

$$\mathbf{h}^{*T} = \begin{bmatrix} 8 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$$

Note that these results agree with those determined in Example 2.4.   ■

If there are repeated roots, then the canonical representation changes form somewhat. For instance, if the root λ_1 has a multiplicity of 3, then (2-16) becomes

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & \cdots \\ 0 & \lambda_1 & 1 & 0 & \cdots \\ 0 & 0 & \lambda_1 & 0 & \cdots \\ 0 & 0 & 0 & \lambda_2 & \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ \vdots \end{bmatrix}u(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2-23}$$

i.e., a *Jordan canonical* $\mathbf{F}$ with all ones along the superdiagonal of the block with λ_1 as diagonal terms. Those superdiagonal terms must be ones in a minimally dimensioned single input–single output system model; they need not be all ones with multiple inputs or outputs.

If some of the eigenvalues are complex conjugate pairs, then the canonical $\mathbf{F}$ matrix will have complex entries along its diagonal, and $\mathbf{x}(t)$ will have some complex components. The *modified canonical form* [2, 5] maintains the desirable characteristic of mode separation, while regaining a totally real-valued system description. Given a canonical system model with $j = \sqrt{-1}$ and

$$
\mathbf{F} = \begin{bmatrix} (\sigma + j\omega) & 0 & | & 0 & \cdots \\ 0 & (\sigma - j\omega) & | & 0 & \cdots \\ \underline{} & \underline{} & & & \\ 0 & 0 & & \lambda_3 & \\ \vdots & \vdots & & & \ddots \end{bmatrix} \tag{2-24}
$$

A similarity transformation described by

$$
\mathbf{T} = \begin{bmatrix} 1/2 & -j/2 & | & 0 & \cdots \\ 1/2 & j/2 & | & 0 & \cdots \\ \underline{} & \underline{} & & & \\ 0 & 0 & & 1 & \\ \vdots & \vdots & & & \ddots \end{bmatrix}, \qquad \mathbf{T}^{-1} = \begin{bmatrix} 1 & 1 & | & 0 & \cdots \\ j & -j & | & 0 & \cdots \\ \underline{} & \underline{} & & & \\ 0 & 0 & & 1 & \\ \vdots & \vdots & & & \ddots \end{bmatrix} \tag{2-25}
$$

yields an equivalent real-valued system description as in (2-8) to (2-10), with

$$
\mathbf{F}^* = \begin{bmatrix} \sigma & \omega & | & 0 & \cdots \\ -\omega & \sigma & | & 0 & \cdots \\ \underline{} & \underline{} & & & \\ 0 & 0 & & \lambda_3 & \\ \vdots & \vdots & & & \ddots \end{bmatrix} \tag{2-26}
$$

This idea can be generalized to the *modified Jordan canonical form* as well (see Brockett [2]).

Note that the entire previous discussion assumed that $n > p$ in the differential equation (2-1) or transfer function (2-3), i.e., that the order of the denominator of $G(s)$ is greater than the order of the numerator. If $n = p$, then the equivalent state space model is of a generalized form:

$$
\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{b}u(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2-27}
$$

$$
z(t) = \mathbf{h}^{\mathsf{T}}\mathbf{x}(t) + du(t) \tag{2-28}
$$

where now the output involves a direct feedthrough of the input u. The corresponding generalization of Eq. (2-6) is

$$
G(s) = \mathbf{h}^{\mathsf{T}}[s\mathbf{I} - \mathbf{F}]^{-1}\mathbf{b} + d \tag{2-29}
$$

EXAMPLE 2.6   Consider a system described by the "lead-lag" network, $y(s) = G(s)u(s)$, with

$$G(s) = \frac{s + a}{s + b}$$

This is equivalent to

$$G(s) = \frac{s + b + a - b}{s + b} = 1 + \frac{a - b}{s + b}$$

The term $(a - b)/(s + b)$ can be represented equivalently in standard controllable form

$$\dot{x}(t) = -bx(t) + u(t)$$
$$z'(t) = [a - b]x(t)$$

and then

$$z(t) = z'(t) + u(t) = [a - b]x(t) + u(t)$$

Figure 2.7 presents two equivalent block diagrams for this representation. It is obvious that (a) represents the equations just written. Diagram (b) obeys the same state differential equation, and then

$$z(t) = ax(t) + \dot{x}(t) = ax(t) - bx(t) + u(t) = [a - b]x(t) + u(t) \qquad \blacksquare$$
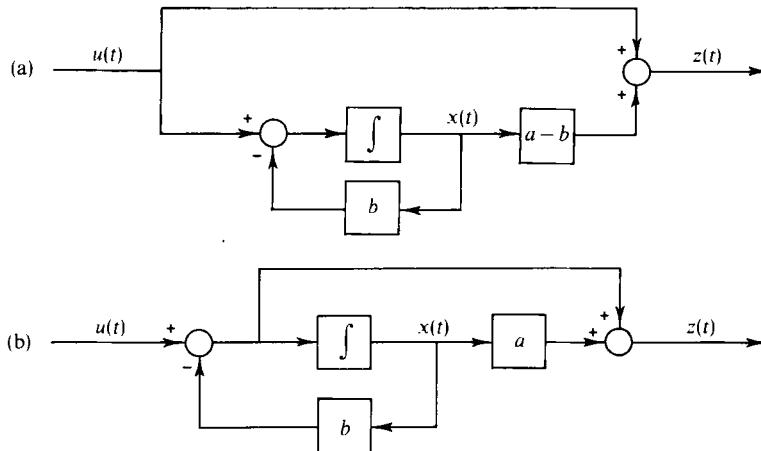
. FIG. 2.7   State space models of $G(s) = (s + a)/(s + b)$.

State space representations are readily extended to *multiple input–multiple output* systems. Assume that there are r inputs into and m outputs from a system, described by r-dimensional $u(t)$ and m-dimensional $z(t)$, respectively. Then the state space model of such a time-invariant system would be

$$\dot{\mathbf{x}}(t) = \mathbf{Fx}(t) + \mathbf{Bu}(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2-30}$$

$$\mathbf{z}(t) = \mathbf{Hx}(t) \tag{2-31}$$

where $\mathbf{F}$ is the unaltered n-by-n matrix describing the homogeneous system dynamics, $\mathbf{B}$ is an n-by-r input matrix, and $\mathbf{H}$ is an m-by-n output matrix. Note the convention on $\mathbf{H}$: if the output is scalar, then $\mathbf{H}$ is a 1-by-n matrix, but this can be viewed as the transpose of an n-dimensional column vector, hence the previous notation $\mathbf{h}^T$. Analogous to the previous discussion, the output equation can be generalized to

$$\mathbf{z}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \qquad (2\text{-}32)$$

but this direct feedthrough structure is often not required.

An equivalent time-invariant multiple input–multiple output model can be developed in the form of a matrix transfer function, whose entries would be the transfer functions of the individual components of the input and output vectors:

$$\mathbf{z}(s) = \mathbf{G}(s)\mathbf{u}(s) \qquad (2\text{-}33)$$

$$\mathbf{G}(s) = \begin{bmatrix} G_{11}(s) & \cdots & G_{1r}(s) \\ \vdots & & \vdots \\ G_{m1}(s) & \cdots & G_{mr}(s) \end{bmatrix} = \mathbf{H}[s\mathbf{I} - \mathbf{F}]^{-1}\mathbf{B} + \mathbf{D} \qquad (2\text{-}34)$$

Again, $\mathbf{D}$ is often $\mathbf{0}$. However, this model is usually rather cumbersome, and the state space model is preferable.

Time-varying system models are generated readily through state space methods—the matrices defining the system structure simply vary with time:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \qquad (2\text{-}35)$$

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) \qquad (2\text{-}36)$$

or, generalized to

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \qquad (2\text{-}37)$$

Laplace transform methods are not readily extended to these cases. Such time-varying linear models arise most naturally from perturbations of a nonlinear set of relations about a nominal solution to the original nonlinear equations. This will be discussed further once we establish conditions under which the existence of such a nominal solution can be assumed.

The relations just given serve to define the most general deterministic linear system model. In Chapter 4, these will be extended to the stochastic linear system model of

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \qquad (2\text{-}38)$$

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}(t) \qquad (2\text{-}39)$$

where $\mathbf{w}(t)$ is a dynamic driving noise and $\mathbf{v}(t)$ is a measurement corruption noise.

A *nonlinear state model* of a system can be described through a state differential equation and output relation of

$$\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), \mathbf{u}(t), t]; \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \qquad\qquad (2\text{-}40)$$

$$\mathbf{z}(t) = \mathbf{h}[\mathbf{x}(t), \mathbf{u}(t), t] \qquad\qquad\qquad (2\text{-}41)$$

where $\mathbf{f}[\cdot, \cdot, \cdot]$ is a mapping from $R^n \times R^r \times R^1$ into R^n [given any $\mathbf{x}(t) \in R^n$, $\mathbf{u}(t) \in R^r$, and $t \in R^1$ (= the real line), $\mathbf{f}$ can be evaluated to yield a vector $\dot{\mathbf{x}}(t) \in R^n$] and $\mathbf{h}[\cdot, \cdot, \cdot]$ is a mapping from $R^n \times R^r \times R^1$ into R^m. For time-invariant nonlinear models, $\mathbf{f}$ and $\mathbf{h}$ are not explicit functions of time, and analogous to the previous discussion for linear models, $\mathbf{h}$ may not be an explicit function of $\mathbf{u}(t)$.

EXAMPLE 2.7   A model of a satellite in planar orbit can be established through the approximation of a unit point mass in an inverse square law force field. Let r be the range from the force field center (earth center) to the satellite and θ be the angle between a reference coordinate axis through the field center and the line from the center to the satellite. Assume the satellite can thrust radially with thrust u_r and tangentially with thrust u_t. The motion of the satellite is then governed by a pair of coupled second order equations:

$$\ddot{r}(t) = r(t)\dot{\theta}^2(t) - \frac{G}{r^2(t)} + u_r(t)$$

$$\ddot{\theta}(t) = -\frac{2}{r(t)}\dot{\theta}(t)\dot{r}(t) + \frac{1}{r(t)} u_t(t)$$

These relations can be put into the form of (2-40) by using the states $x_1 = r, x_2 = \dot{r}, x_3 = \theta,$ and $x_4 = \dot{\theta}$ to write

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \end{bmatrix} = \begin{bmatrix} f_1[\mathbf{x}(t), \mathbf{u}(t), t] \\ f_2[\mathbf{x}(t), \mathbf{u}(t), t] \\ f_3[\mathbf{x}(t), \mathbf{u}(t), t] \\ f_4[\mathbf{x}(t), \mathbf{u}(t), t] \end{bmatrix} = \begin{bmatrix} x_2(t) \\ x_1(t)x_4^2(t) - \dfrac{G}{x_1^2(t)} + u_r(t) \\ x_4(t) \\ -\dfrac{2}{x_1(t)} x_4(t)x_2(t) + \dfrac{1}{x_1(t)} u_t(t) \end{bmatrix} \blacksquare$$

Equations (2-40) and (2-41) are the form of a general deterministic nonlinear state-described system model, and Chapter 11 (Volume 2) will extend them to the stochastic model case. A more general class of system models called *dynamic system models* can be defined (see Desoer [7], etc.), but this level of generality will be adequate for our purposes.

2.3   SOLUTIONS TO STATE DIFFERENTIAL EQUATIONS

In this section, the solution to the state differential equations just described will be presented, starting with the most general case and then considering the simplifications made possible by progressively more restrictive models [4, 13].

To describe some underlying assumptions simply, consider the homogeneous nonlinear differential equation

$$\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), t]; \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2-42}$$

First, assume that $\mathbf{f}(\cdot, \cdot)$ is piecewise continuous in its second argument. In other words, for each $\mathbf{x}_i \in R^n$, the mapping $\mathbf{f}(\mathbf{x}_i, \cdot)$ is continuous except possibly at a finite number of points, where left and right limits are well defined. Next, assume that $\mathbf{f}(\cdot, \cdot)$ is Lipschitz in its first argument, which is to say that there exists a piecewise continuous function $k(\cdot)$ such that, for all $t \in [0, \infty)$ and all $\mathbf{x}_1, \mathbf{x}_2 \in R^n$,

$$\left\| \mathbf{f}(\mathbf{x}_1, t) - \mathbf{f}(\mathbf{x}_2, t) \right\| < k(t) \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\| \tag{2-43}$$

where $\|\mathbf{v}\| = \max_i |v_i|$ for $\mathbf{v} \in R^n$. These two assumptions together imply, for *any* function $\psi(\cdot)$ mapping $[t_0, \infty)$ into R^n, that the function that maps t into $\mathbf{f}[\psi(t), t]$ is a piecewise continuous function. Therefore, for any such $\psi(\cdot)$ we can integrate $\mathbf{f}[\psi(t), t]$ with respect to time, and then the function that maps t into $\int_0^t \mathbf{f}[\psi(\tau), \tau] \, d\tau$ is continuous. Moreover, by the fundamental theorem of calculus, its derivative is equal to $\mathbf{f}[\psi(t), t]$ for all $t \in [t_0, \infty)$ except at the possible points of discontinuity.

With this introduction, it is possible to state one form of the *fundamental theorem of differential equations*: consider the differential equation and initial condition

$$\dot{\mathbf{x}}(t) = \mathbf{f}[\mathbf{x}(t), \mathbf{u}(t), t]; \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2-44}$$

where the function $\mathbf{f}(\cdot, \cdot, \cdot)$ that maps $R^n \times R^r \times [t_0, \infty)$ into R^n is assumed to be

(1) Lipschitz in its first argument,
(2) continuous in its second argument, and
(3) piecewise continuous in its third argument.

Then, for each $\mathbf{x}_0 \in R^n$ and each $t_0 \in [0, \infty)$ and any piecewise continuous r-vector-valued function $\mathbf{u}(\cdot)$, there *exists* a *unique continuous* mapping $\phi(\cdot)$ from $[0, \infty)$ into R^n such that

$$\phi(t_0) = \mathbf{x}_0 \tag{2-45}$$

and

$$\dot{\phi}(t) = \mathbf{f}[\phi(t), \mathbf{u}(t), t] \tag{2-46}$$

for all $t \in [0, \infty)$ except at the possible points of discontinuity. The function $\phi(\cdot)$ is called the *solution* to the differential equation (2-44), and its value depends only on t, t_0, $\mathbf{x}_0$, and the values that $\mathbf{u}$ assumes in the interval $[t_0, t]$.

The proof will not be presented here (see Desoer [7], for example), but is based upon establishing local existence, using successive approximations to

demonstrate constructively this existence globally, and then proving uniqueness. This procedure motivates the proof of existence of solutions to nonlinear stochastic differential equations as well, to be discussed in Chapter 11 (Volume 2).

Once a nominal solution to a nonlinear differential equation can be found, perturbations about this nominal solution can be considered. For a given x_0, t_0, and input function $u_0(\cdot)$, let (2-44) have a solution denoted as $x_0(\cdot)$ for $t \in [t_0, \infty)$. What happens if the initial condition were perturbed to $(x_0 + \Delta x_0)$ and/or the input were perturbed to $[u_0(\cdot) + \Delta u(\cdot)]$? If these are "small" perturbations, we would expect the perturbed solution to be $[x_0(\cdot) + \Delta x(\cdot)]$ with $\Delta x(t)$ "small" for all $t \in [t_0, \infty)$. (More precise conditions can be stated; see Desoer and Wong [8]), so that Taylor series about the nominal could be exploited:

$$\dot{x}(t) = f[x(t), u(t), t]$$
$$= f[x_0(t), u_0(t), t] + F(t)\{x(t) - x_0(t)\} + B(t)\{u(t) - u_0(t)\} + \cdots \quad (2\text{-}47)$$

where

$$F(t) = \frac{\partial f}{\partial x}\bigg|_{x_0(t),\, u_0(t),\, t} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}\Bigg|_{x_0(t),\, u_0(t),\, t} \quad (2\text{-}48)$$

$$B(t) = \frac{\partial f}{\partial u}\bigg|_{x_0(t),\, u_0(t),\, t} = \begin{bmatrix} \dfrac{\partial f_1}{\partial u_1} & \cdots & \dfrac{\partial f_1}{\partial u_r} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial u_1} & \cdots & \dfrac{\partial f_n}{\partial u_r} \end{bmatrix}\Bigg|_{x_0(t),\, u_0(t),\, t} \quad (2\text{-}49)$$

Since $\dot{x}_0(t) = f[x_0(t), u_0(t), t]$, (2-47) yields

$$d\{x(t) - x_0(t)\}/dt = F(t)\{x(t) - x_0(t)\} + B(t)\{u(t) - u_0(t)\} + \cdots$$

By neglecting the higher order terms, one obtains an approximation to the true differential equation satisfied by $\{x(t) - x_0(t)\}$, called the *linearized perturbation equation* or equation of the first variation [2, 7], in the general form of a time-varying linear differential equation:

$$\dot{\delta x}(t) = F(t)\,\delta x(t) + B(t)\,\delta u(t) \quad (2\text{-}50)$$

where $\delta x(t) \cong \{x(t) - x_0(t)\}$ and $\delta u(t) = \{u(t) - u_0(t)\}$.

EXAMPLE 2.8   Return to the model of satellite motion discussed in Example 2.7. Perturbations about a nominal trajectory can be described approximately by the model

$$\dot{\delta}\mathbf{x}(t) = \mathbf{F}(t)\,\delta\mathbf{x}(t) + \mathbf{B}(t)\,\delta\mathbf{u}(t)$$

where

$$\mathbf{F}(t) = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}_0(t),\,\mathbf{u}_0(t),\,t} = \left.\begin{bmatrix} 0 & 1 & 0 & 0 \\ \left[x_4{}^2 + \dfrac{2G}{x_1{}^3}\right] & 0 & 0 & 2x_1 x_4 \\ 0 & 0 & 0 & 1 \\ \left[\dfrac{2x_2 x_4}{x_1{}^2} - \dfrac{u_t}{x_1{}^2}\right] & -\dfrac{2x_4}{x_1} & 0 & -\dfrac{2x_2}{x_1} \end{bmatrix}\right|_{\mathbf{x}_0(t),\,\mathbf{u}_0(t),\,t}$$

$$\mathbf{B}(t) = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right|_{\mathbf{x}_0(t),\,\mathbf{u}_0(t),\,t} = \left.\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/x_1 \end{bmatrix}\right|_{\mathbf{x}_0(t),\,\mathbf{u}_0(t),\,t}$$

One particular solution admitted by the original nonlinear equation is that of a circular orbit: $r(t) = r_0$, $\dot{r}(t) = 0$, $\theta(t) = \omega t$, $\dot{\theta}(t) = \omega$, $u_t(t) = u_r(t) = 0$, $G = r_0{}^3\omega^2$ for all time. For this nominal, $\mathbf{F}$ and $\mathbf{B}$ are in fact time invariant:

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2r_0\omega \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega/r_0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/r_0 \end{bmatrix} \qquad \blacksquare$$

The solution to linear differential equations can be written explicitly. If a proposed solution form satisfies the differential equation and the initial conditions, then it is *the unique solution* because the assumptions of the previous theorem will be met whenever $[\mathbf{B}(\cdot)\mathbf{u}(\cdot)]$ is piecewise continuous.

The solution to the linear time-varying differential equation and initial condition

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \qquad (2\text{-}51)$$

for $\mathbf{F}(\cdot)$ and $[\mathbf{B}(\cdot)\mathbf{u}(\cdot)]$ piecewise continuous is given by

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \qquad (2\text{-}52)$$

where $\mathbf{\Phi}(\cdot, \cdot)$ is the *state transition matrix* defined as the n-by-n matrix that satisfies the differential equation and initial condition

$$d[\mathbf{\Phi}(t, t_0)]/dt = \mathbf{F}(t)\mathbf{\Phi}(t, t_0) \qquad (2\text{-}53a)$$

$$\mathbf{\Phi}(t_0, t_0) = \mathbf{I} \qquad (2\text{-}53b)$$

Significant properties of the state transition matrix include:

(1) $\boldsymbol{\Phi}(t, t_0)$ is *uniquely* defined for all t and t_0 in $[0, \infty)$.

(2)   The state transition matrix to propagate from any t_1 to t_3 equals the product of the separate transition matrices from t_1 to t_2 and t_2 to t_3 (called the semigroup property):

$$\boldsymbol{\Phi}(t_3, t_1) = \boldsymbol{\Phi}(t_3, t_2)\boldsymbol{\Phi}(t_2, t_1) \tag{2-54}$$

(3) $\boldsymbol{\Phi}(t, t_0)$ is nonsingular (invertible) and

$$\boldsymbol{\Phi}(t, t_0)\boldsymbol{\Phi}(t_0, t) = \boldsymbol{\Phi}(t, t) = \mathbf{I}$$

so that

$$\boldsymbol{\Phi}^{-1}(t, t_0) = \boldsymbol{\Phi}(t_0, t) \tag{2-55}$$

Let us show that the proposed solution, (2-52), does in fact satisfy both the differential equation and initial condition in (2-51):

$$\mathbf{x}(t_0) = \boldsymbol{\Phi}(t_0, t_0)\mathbf{x}_0 + \int_{t_0}^{t_0} \boldsymbol{\Phi}(t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

$$= \mathbf{I}\mathbf{x}_0 + \mathbf{0} = \mathbf{x}_0$$

To demonstrate similar satisfaction of the differential equation will require use of Leibnitz' rule:

$$\frac{d}{dt} \int_{A(t)}^{B(t)} \mathbf{f}(t, \tau)\, d\tau = \int_{A(t)}^{B(t)} \frac{\partial \mathbf{f}(t, \tau)}{\partial t}\, d\tau + \mathbf{f}[t, B(t)]\frac{dB}{dt} - \mathbf{f}[t, A(t)]\frac{dA}{dt} \tag{2-56}$$

Differentiating (2-52) thus yields

$$\frac{d}{dt}\mathbf{x}(t) = \dot{\boldsymbol{\Phi}}(t, t_0)\mathbf{x}_0 + \boldsymbol{\Phi}(t, t)\mathbf{B}(t)\mathbf{u}(t) + \int_{t_0}^{t} \dot{\boldsymbol{\Phi}}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

$$= \mathbf{F}(t)\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \mathbf{B}(t)\mathbf{u}(t) + \int_{t_0}^{t} \mathbf{F}(t)\boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

$$= \mathbf{F}(t)\left[\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \right] + \mathbf{B}(t)\mathbf{u}(t)$$

But the terms in brackets are just the assumed form for $\mathbf{x}(t)$, so this is the desired solution.

Equation (2-52) is the appropriate solution form for linear time-invariant differential equations as well, but the state transition matrix in this case can be characterized further. If $\mathbf{F}$ is a constant matrix, then the associated $\boldsymbol{\Phi}(t, t_0)$ is not a function of the separate arguments t and t_0 but is a function only of the single parameter $(t - t_0)$. Thus, the general defining relationship for the state transition matrix reduces to

$$d[\boldsymbol{\Phi}(t - t_0)]/dt = \mathbf{F}\boldsymbol{\Phi}(t - t_0); \qquad \boldsymbol{\Phi}(0) = \mathbf{I} \tag{2-57}$$

to which the solution can be expressed as the matrix exponential

$$\boldsymbol{\Phi}(t, t_0) = \boldsymbol{\Phi}(t - t_0) = e^{\mathbf{F}(t - t_0)} \qquad (2\text{-}58a)$$

Another expression for $\boldsymbol{\Phi}(t - t_0)$ for time-invariant systems can be obtained by taking the Laplace transform of (2-57), letting $t_0 = 0$:

$$s\boldsymbol{\Phi}(s) - \boldsymbol{\Phi}(t - t_0 = 0) = \mathbf{F}\boldsymbol{\Phi}(s)$$
$$s\boldsymbol{\Phi}(s) - \mathbf{F}\boldsymbol{\Phi}(s) = \boldsymbol{\Phi}(t - t_0 = 0)$$
$$[s\mathbf{I} - \mathbf{F}]\boldsymbol{\Phi}(s) = \mathbf{I}$$
$$\boldsymbol{\Phi}(s) = [s\mathbf{I} - \mathbf{F}]^{-1} \qquad (2\text{-}58b)$$

Thus $\boldsymbol{\Phi}(t - t_0)$ is the inverse Laplace transform of $[s\mathbf{I} - \mathbf{F}]^{-1}$, the resolvent matrix mentioned previously.

2.4 DISCRETE-TIME MEASUREMENTS

In many applications of estimation or control theory to actual problems, a digital computer performs online computations using data samples from a continuous-time dynamic process, generally (but not necessarily) taken at a fixed sample rate. Consequently, a discrete-time measurement equation will often be more pertinent than the continuous-time output equations already described. If t_i is a measurement sample time, then the measurement data can be represented as the sampled versions of (2-41) in the nonlinear case:

$$\mathbf{z}(t_i) = \mathbf{h}[\mathbf{x}(t_i), \mathbf{u}(t_i), t_i] \qquad (2\text{-}59)$$

or of (2-36) or (2-37) for the linear case

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) \qquad (2\text{-}60a)$$

or

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{D}(t_i)\mathbf{u}(t_i) \qquad (2\text{-}60b)$$

It will be seen later that the computer software for implementing an optimal estimator or controller will embody a discrete-time model that is "equivalent" to the continuous-time model, in the sense that the discrete-time model's values of $\mathbf{x}(t_1)$, $\mathbf{x}(t_2)$, ..., are identical to those of the continuous-time model at these particular times. From Eq. (2-52), we can write for the linear model case:

$$\mathbf{x}(t_{i+1}) = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \int_{t_i}^{t_{i+1}} \boldsymbol{\Phi}(t_{i+1}, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \qquad (2\text{-}61)$$

Since a digital computer is assumed to apply the control, a very typical form of $\mathbf{u}(\cdot)$ would be a piecewise constant function: a measurement sample would be taken, the information processed, and a control input created and held

constant until the following sample time. If we assume

$$\mathbf{u}(t) = \mathbf{u}(t_i) \qquad \text{for all} \quad t \in [t_i, t_{i+1}) \qquad (2\text{-}62)$$

then (2-61) can be written as

$$\mathbf{x}(t_{i+1}) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \left[\int_{t_i}^{t_{i+1}} \mathbf{\Phi}(t_{i+1}, \tau)\mathbf{B}(\tau)\, d\tau \right] \mathbf{u}(t_i) \qquad (2\text{-}63)$$

This and (2-60) are then in the form of a discrete-time difference equation system model

$$\mathbf{x}(i + 1) = \mathbf{\Phi}(i + 1, i)\mathbf{x}(i) + \mathbf{B}(i)\mathbf{u}(i) \qquad (2\text{-}64\text{a})$$

$$\mathbf{z}(i) = \mathbf{H}(i)\mathbf{x}(i) + \mathbf{D}(i)\mathbf{u}(i) \qquad (2\text{-}64\text{b})$$

where i denotes instant, associated here with time t_i. System models of the form of (2-64) sometimes arise naturally from a basic problem application, as well as from "discretizing" a continuous-time model. However, in this case, one is not assured of such properties as nonsingularity of $\mathbf{\Phi}(i + 1, i)$. We will usually be concerned with problems described most fundamentally through differential, as opposed to difference, equations. For these problems, (2-60) and (2-63) will represent the "equivalent discrete-time system model." In the nonlinear case, the discrete-time dynamic model is

$$\mathbf{x}(i + 1) = \boldsymbol{\phi}[\mathbf{x}(i), \mathbf{u}(i), i] \qquad (2\text{-}65)$$

but an explicit "equivalency" relationship as in (2-63) cannot be written in a general context.

2.5 CONTROLLABILITY AND OBSERVABILITY

Observability and controllability [2, 9–12, 14, 16, 18] are properties of a specific state space representation for a system, rather than of the system itself. Thus, certain state space models will be more suitable for estimation or control purposes than others, even though both might accurately portray the input–output characteristics of a system.

Controllability is concerned with the effect of inputs upon states of a system model. A continuous-time system representation is said to be *completely controllable* if, for any vectors $\mathbf{x}_0$, $\mathbf{x}_1 \in R^n$ and any time t_0, there exists a piecewise continuous control function $\mathbf{u}(\cdot)$ such that the solution of the describing differential equation with $\mathbf{x}(t_0) = \mathbf{x}_0$ satisfies $\mathbf{x}(t_1) = \mathbf{x}_1$ for some finite t_1. In other words, a system model is completely controllable if any and all initial state variables $x_i(t_0) = x_{0i}$ can be transferred to any final state x_{1i} in finite time by applying a control $\mathbf{u}(t)$, $t_0 \leq t \leq t_1$. Consequently, in order to be completely controllable, the system model structure must at least be such that $\mathbf{u}$ can affect all of the state variables. It is possible to talk of a system being "controllable

at a given t_0" (rather than any t_0), "controllable from $(\mathbf{x}_0, t_0)$ to $(\mathbf{x}_1, t_1)$" (further specifying particular initial and final conditions), and the like, but complete controllability will be a more significant concept for our purposes.

Consider the linear time-varying model given by Eqs. (2-35)–(2-37). Since the solution of the differential equation (2-35) is

$$\mathbf{x}_1 = \mathbf{x}(t_1) = \mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 + \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \qquad (2\text{-}66)$$

we can premultiply both sides by $\mathbf{\Phi}(t_0, t_1)$ to obtain

$$\mathbf{\Phi}(t_0, t_1)\mathbf{x}_1 = \mathbf{x}_0 + \int_{t_0}^{t_1} \mathbf{\Phi}(t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \qquad (2\text{-}67)$$

But $\mathbf{\Phi}(t_0, t_1)\mathbf{x}_1$ is the result of propagating $\mathbf{x}_1$ backwards in time to time t_0 if no controls are applied. Therefore, complete controllability, being able to reach any $\mathbf{x}_1$ at t_1 from $\mathbf{x}_0$, is equivalent to saying $[\mathbf{\Phi}(t_0, t_1)\mathbf{x}_1 - \mathbf{x}_0]$ can be *any* vector in R^n. This is equivalent to saying the range space of $\int_{t_0}^{t_1} \mathbf{\Phi}(t_0, \tau)\mathbf{B}(\tau) \cdot d\tau$ (the space of possible mapping of piecewise continuous $\mathbf{u}$ by this operator) is all of R^n. It can be shown that this range space is equivalent to the range space of the n-by-n matrix (called the controllability Gramian):

$$\mathbf{W}(t_0, t_1) \triangleq \int_{t_0}^{t_1} \mathbf{\Phi}(t_0, \tau)\mathbf{B}(\tau)\mathbf{B}^T(\tau)\mathbf{\Phi}^T(t_0, \tau)\, d\tau \qquad (2\text{-}68)$$

where the form of $\mathbf{W}(t_0, t_1)$ is motivated by a study of adjoints of linear operators. Thus, the system model (2-35)–(2-37) can be shown to be completely controllable if and only if any of the equivalent criteria are met for some t_1: (1) the range of $\mathbf{W}(t_0, t_1)$ is R^n, (2) the rank of $\mathbf{W}(t_0, t_1)$ is n, (3) $\mathbf{W}(t_0, t_1)$ is nonsingular and thus invertible, (4) $\mathbf{W}(t_0, t_1)$ is positive definite (it is always positive semi-definite), (5) the determinant of $\mathbf{W}(t_0, t_1)$ is nonzero. If the model is completely controllable, then one control which actually transfers $\mathbf{x}(t_0) = \mathbf{x}_0$ to $\mathbf{x}(t_1) = \mathbf{x}_1$ is given by:

$$\mathbf{u}(t) = -\mathbf{B}^T(t)\mathbf{\Phi}^T(t_0, t)\mathbf{W}(t_0, t_1)^{-1}[\mathbf{x}_0 - \mathbf{\Phi}(t_0, t_1)\mathbf{x}_1] \qquad \text{for all} \quad t \in [t_0, t_1]$$
$$(2\text{-}69)$$

If $\mathbf{W}(t_0, t_1)$ is singular and of rank $k < n$, then there are $(n - k)$ "uncontrollable states" in the representation. No matter what control is applied, no influence can be exerted on the system response in those directions of n-space. If it is possible to derive an equivalent model with no uncontrollable states, this would be a preferable basis for controller design. When noise is introduced into the dynamics model, it will be appropriate to investigate controllability with respect to both control inputs and noises, and this will be discussed subsequently.

Although the preceding criteria are precise mathematically, they are difficult to apply in practice because they need only be satisfied for some t_1. This is alleviated somewhat by the fact that the rank of $\mathbf{W}(t_0, t_1)$ is a monotonically

increasing function of t_1, for fixed t_0. However, in the case of a time-invariant linear model, as (2-30)–(2-31), a more practical criterion can be achieved. Here, the range space of the n-by-nr matrix

$$\mathbf{W}_{TI} \triangleq [\mathbf{B} \quad \vdots \quad \mathbf{FB} \quad \vdots \quad \cdots \quad \vdots \quad \mathbf{F}^{n-1}\mathbf{B}] \tag{2-70}$$

can be shown equivalent to the range space of $\mathbf{W}(t_0, t_1)$. Thus, the system model (2-30)–(2-31) is completely controllable if and only if the range space of $\mathbf{W}_{TI}$ is R^n, or its rank is n, or equivalently, if there are n linearly independent columns in $\mathbf{W}_{TI}$. Each column of $\mathbf{W}_{TI}$ represents a vector in state space along which control is possible. If it is possible to control along n linearly independent directions in R^n, i.e., a basis of R^n, it is possible to control in all R^n.

For single input system models, $\mathbf{W}_{TI}$ becomes an n-by-n matrix, so in that case the system model is completely controllable if and only if $\mathbf{W}_{TI}$ is non-singular, i.e., its determinant is nonzero. This can also be shown equivalent to the condition that $[s\mathbf{I} - \mathbf{F}]^{-1}\mathbf{b}$ has no pole-zero cancellations.

Analogously, a discrete-time system model is *completely controllable*, if, for any vectors $\mathbf{x}_0$, $\mathbf{x}_N \in R^n$, there exists a sequence $\mathbf{u}(0), \ldots, \mathbf{u}(N-1)$ such that the solution of the describing difference equation with $\mathbf{x}(0) = \mathbf{x}_0$ satisfies $\mathbf{x}(N) = \mathbf{x}_N$ for some finite N. The linear system representation of (2-64) is completely controllable if and only if the range space of the n-by-n matrix

$$\mathbf{W}_D(0, N) \triangleq \sum_{i=1}^{N} \mathbf{\Phi}(0, i)\mathbf{B}(i-1)\mathbf{B}^T(i-1)\mathbf{\Phi}^T(0, i) \tag{2-71}$$

is all of R^n, or any equivalent statements as after (2-68). The corresponding time-invariant linear system model is completely controllable if and only if the rank of the n-by-nr matrix $\mathbf{W}_{DTI}$ is n, where

$$\mathbf{W}_{DTI} \triangleq [\mathbf{B} \quad \vdots \quad \mathbf{\Phi}\mathbf{B} \quad \vdots \quad \cdots \quad \vdots \quad \mathbf{\Phi}^{n-1}\mathbf{B}] \tag{2-72}$$

On the other hand, observability is concerned with the effect of states of a model upon the outputs. A continuous-time system representation is *completely observable* if, given $\mathbf{z}(t)$ and $\mathbf{u}(t)$ for all $t \in [t_0, t_1]$, it is possible to deduce $\mathbf{x}(t)$ for $t \in [t_0, t_1]$. Thus, a system model is completely observable if *any* state $x_i(t)$ can be determined exactly for $t \in [t_0, t_1]$ from knowledge of only the input and output over the interval $[t_0, t_1]$. To be completely observable, the representation structure must be such that the output $\mathbf{z}(t)$ is affected in some manner by the change of any single state variable. Moreover, the effect of any one state variable on the output must be distinguishable from the effect of any other state variable.

EXAMPLE 2.9 Consider a homogeneous system model as in Fig. 2.8a. Here, both x_1 and x_2 affect the output z, but there would be no way to obtain separate information about x_1 and x_2 just by observing the output z. From observations of z this model would appear identical to that depicted in Fig. 2.8b. ■
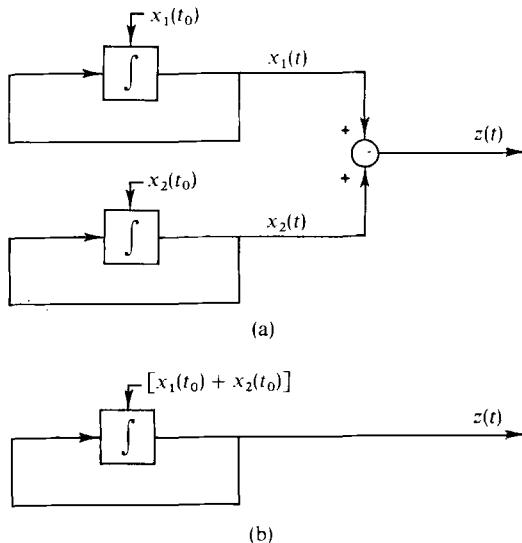
FIG. 2.8   Observability of system models. (a) Original model. (b) Equivalent model.

When a complete system model is generated by combining separate component models, it is not uncommon for the result to involve unobservable states. For instance, in a redundant system, a state to model a bias in one instrument might be indistinguishable from a similar bias state in a redundant instrument, considering only the effects on the output. If an estimator were based upon such a system model, the estimation errors along certain directions of state space would not decrease, regardless of how long measurements were taken. In such cases, it would be appropriate to combine such different physical quantities into a single state variable, such as to let a state be the sum of the biases in the redundant system example, to achieve an observable system model.

Now let us ask whether or not the linear time-varying system model described by (2-35) and (2-37) is completely observable. By the form of the solution to the differential equation, it is necessary and sufficient to be able to deduce $\mathbf{x}(t_0)$ from knowledge of $\mathbf{u}(t)$ and $\mathbf{z}(t)$ for all $t \in [t_0, t_1]$. Using the solution we can write

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t)$$

$$= \mathbf{H}(t)\left[\mathbf{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \right] + \mathbf{D}(t)\mathbf{u}(t)$$

so that

$$\mathbf{H}(t)\mathbf{\Phi}(t, t_0)\mathbf{x}_0 = \mathbf{z}(t) - \mathbf{D}(t)\mathbf{u}(t) - \int_{t_0}^{t} \mathbf{H}(t)\mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \qquad (2\text{-}73)$$

Thus we are asking, under what conditions can this equation be solved for $\mathbf{x}_0$ uniquely? This is equivalent to asking, is the null space of $\mathbf{H}(t)\mathbf{\Phi}(t, t_0) \cdot$, the set of all vectors $\mathbf{x} \in R^n$ that are mapped into the zero function over the interval $[t_0, t_1]$, restricted to $\mathbf{0} \in R^n$? This is a difficult question to answer, but it can be shown that the null space of $\mathbf{H}(t)\mathbf{\Phi}(t, t_0) \cdot$ is the same as the null space of the n-by-n matrix (the observability Gramian)

$$\mathbf{M}(t_0, t_1) \triangleq \int_{t_0}^{t_1} \mathbf{\Phi}^T(\tau, t_0)\mathbf{H}^T(\tau)\mathbf{H}(\tau)\mathbf{\Phi}(\tau, t_0)\, d\tau \tag{2-74}$$

As in the case of controllability, the form of $\mathbf{M}(t_0, t_1)$ is motivated by adjoints of linear operators. Consequently, the model (2-35)–(2-37) is completely observable if and only if any of the following criteria are met for some t_1: (1) the null space of $\mathbf{M}(t_0, t_1)$ is $\mathbf{0} \in R^n$, (2) $\mathbf{M}(t_0, t_1)$ is nonsingular, i.e., invertible, (3) $\mathbf{M}(t_0, t_1)$ is positive definite, (4) the determinant of $\mathbf{M}(t_0, t_1)$ is nonzero. If $\mathbf{M}(t_0, t_1)$ is of rank $k < n$, then it is said that there are $(n - k)$ "unobservable states" in the model. If the model is completely observable, then $\mathbf{x}_0$ can be determined uniquely from $\mathbf{u}(t)$ and $\mathbf{z}(t)$, $t_0 \le t \le t_1$, by

$$\mathbf{x}_0 = \mathbf{M}(t_0, t_1)^{-1} \int_{t_0}^{t_1} \mathbf{\Phi}^T(\tau, t_0)\mathbf{H}^T(\tau)\mathbf{v}(\tau)\, d\tau \tag{2-75}$$

where

$$\mathbf{v}(\tau) = \mathbf{z}(\tau) - \mathbf{D}(\tau)\mathbf{u}(\tau) - \int_{t_0}^{\tau} \mathbf{H}(\tau)\mathbf{\Phi}(\tau, \sigma)\mathbf{B}(\sigma)\mathbf{u}(\sigma)\, d\sigma \tag{2-76}$$

Using that value of $\mathbf{x}_0$, the entire state history over $[t_0, t_1]$ can be determined from

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{\tau} \mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \tag{2-77}$$

As before, a more practical test is possible for time-invariant linear models. The system representation given by (2-30)–(2-31) is completely observable if and only if the range space of $\mathbf{M}_{\mathrm{TI}}$ is R^n, where $\mathbf{M}_{\mathrm{TI}}$ is the n-by-nm matrix

$$\mathbf{M}_{\mathrm{TI}} \triangleq [\mathbf{H}^T \quad\vdots\quad \mathbf{F}^T\mathbf{H}^T \quad\vdots\quad \cdots \quad\vdots\quad (\mathbf{F}^T)^{n-1}\mathbf{H}^T] \tag{2-78}$$

or equivalently, if its rank is n or if there are n linearly independent columns in $\mathbf{M}_{\mathrm{TI}}$ (again, if we can observe along a basis of R^n, then we can observe any vector in R^n).

For single output systems, the preceding criterion becomes the condition that the n-by-n $\mathbf{M}_{\mathrm{TI}}$ is nonsingular, with nonzero determinant. This is equivalent to the condition that $\mathbf{h}^T[s\mathbf{I} - \mathbf{F}]^{-1}$ has no pole-zero cancellations.

The corresponding discrete-time result is that the model described by (2-64) is completely observable if and only if the null space of the n-by-n matrix

$$\mathbf{M}_{\mathrm{D}}(0, N) \triangleq \sum_{i=1}^{N} \mathbf{\Phi}^T(i, 0)\mathbf{H}^T(i)\mathbf{H}(i)\mathbf{\Phi}(i, 0) \tag{2-79}$$

is $\mathbf{0} \in R^n$ for some finite N, or any equivalent statements as after (2-74). Note that the rank of each term in the sum is at most m, so there is in general some minimal number $N \geq n/m$ of measurements that must be taken before the model can be completely observable. A corresponding time-invariant linear model is completely observable if and only if the rank of the n-by-nm matrix $\mathbf{M}_{\text{DTI}}$ is n, where

$$\mathbf{M}_{\text{DTI}} \triangleq [\mathbf{H}^T \quad \vdots \quad \mathbf{\Phi}^T \mathbf{H}^T \quad \vdots \quad \cdots \quad \vdots \quad (\mathbf{\Phi}^T)^{n-1} \mathbf{H}^T] \qquad (2\text{-}80)$$

There is a noteworthy resemblance between the observability and controllability results just discussed. The interrelationship is a *duality* relationship that can be exploited substantially in linear system theory, but we will not pursue this matter further at this point.

2.6 SUMMARY

The state differential equation (2-35) and discrete-time output equation (2-60) will provide the basic linear deterministic system model, the structure of which will be extended in ensuing chapters by adding uncertainty to both the dynamics and measurement relations. The dynamic system response can be characterized by solving the state differential equation, facilitated by means of the concept of the state transition matrix. For such analysis, the interrelationships among inputs, states, and outputs are properly specified through the concepts of controllability and observability. Given a particular system model, an infinite variety of related time-domain models can represent the same input–output characteristics. Some particular forms are especially useful, in that they separate system modes or yield minimum required computations to depict the system behavior, and these can be exploited in practical applications.

In a similar manner, Eqs. (2-40) and (2-59) define a general nonlinear deterministic system model. The existence of solutions to nonlinear differential equations can be established under rather nonrestrictive assumptions, but their form cannot be characterized as fully as in the linear case. This model structure will be extended in Chapter 11 (Volume 2) to the stochastic case, motivated by the insights gained from the simpler models discussed in this chapter.

REFERENCES

1. Athans, M., "The Relationship of Alternate State-Space Representations in Linear Filtering Problems," *IEEE Trans. Automatic Control* **AC 12**, 775–776 (1967).
2. Brockett, R. W., *Finite Dimensional Linear Systems*. Wiley, New York, 1970.
3. Chen, C. T., *Introduction to Linear System Theory*. Holt, New York, 1970.
4. Coddington, E. A., and Levinson, N., *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.
5. D'Azzo, J. J., and Houpis, C. H., *Linear Control Systems Analysis and Design: Conventional and Modern*. McGraw-Hill, New York, 1975.
6. DeRusso, P. M., Roy, R. J., and Close, C. M., *State Variables for Engineers*. Wiley, New York, 1965.

7. Desoer, C. A., *Notes for a Second Course on Linear Systems.* Van Nostrand-Reinhold, Princeton, New Jersey, 1970.

8. Desoer, C. A., and Wong, K. K., "Small Signal Behavior of Nonlinear Lumped Networks," *Proc. IEEE* **56**, 14–22 (1968).

9. Kalman, R. E., "Contributions to the Theory of Optimal Control," *Bol. Soc. Mat. Mexicana Ser. 2* **5**, 102 (1960).

10. Kalman, R. E., "Mathematical Description of Linear Dynamical Systems," *J. SIAM, Ser. A* **1**, 152 (1963).

11. Kalman, R. E., Englar, T. S., and Bucy, R. S., *Fundamental Study of Adaptive Control Systems*, ASD-TR-61-27, Wright-Patterson AFB, Ohio, 1961.

12. Kalman, R. E., Ho, Y. C., and Narendra, K. S., "Controllability of Linear Dynamical Systems," in *Contributions to Differential Equations*, Vol. 1, pp. 189–213. Wiley, New York, 1962.

13. Kaplan, W., *Ordinary Differential Equations.* Addison-Wesley, Reading, Massachusetts, 1960.

14. Kreindler, E., and Sarachik, P. E., "On the Concepts of Controllability and Observability of Linear Systems," *IEEE Trans. Automatic Control* **AC 9**, 129–136 (1964).

15. Ogata, K., *State Space Analysis of Control Systems.* Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

16. Porter, W. A., *Modern Foundations of Systems Engineering.* Macmillan, New York, 1966.

17. Schultz, D. G., and Melsa, J. L., *State Functions and Linear Control Systems.* McGraw-Hill, New York, 1967.

18. Sorenson, H. W., "Controllability and Observability of Linear Stochastic Time-Discrete Control Systems," in *Advances in Control Systems* (C. T. Leondes, ed.), Vol. VI. Academic Press, New York, 1969.

19. Zadeh, L. A., and Desoer, C. A., *Linear System Theory.* McGraw-Hill, New York, 1963.

PROBLEMS

2.1 Suppose we have a single input–single output system described by the following transfer function:

$$G(s) = (s^3 + 3s^2 + 5s + 8)/(s^4 + 7s^3 + 14s^2 + 8s)$$

(a) Derive the standard observable and standard controllable phase variable forms. Note that the $\mathbf{F}$ matrix of $\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{b}u(t)$ is singular—how is this interpreted physically? Draw the block diagrams for both and note the pattern of the feedback and feedforward loops.

(b) Represent the system in terms of canonical variables; draw the block diagram.

(c) Are the system representations just obtained observable and controllable?

2.2 For a system modeled by transfer function of

$$G(s) = [10(s + 4)]/(s^3 + 3s^2 + 2s)$$

(a) evaluate the standard controllable form of state variable representation,

(b) obtain the canonical form state variable representation from the phase variables, with and without the Vandermonde matrix, and

(c) obtain the canonical representation directly from the transfer function.

2.3 For a system modeled by $G(s) = 1/[(s^2 + 6s + 25)(s + 1)]$ generate the canonical form state variable model involving complex $\mathbf{F}$ and $\mathbf{x}(t)$. Transform the result appropriately to make $\mathbf{F}$ and $\mathbf{x}(t)$ be composed of only real-valued terms. From the transfer function, write the standard controllable form, and then obtain the same form as just stated without going through the intermediate step of complex $\mathbf{F}$ and $\mathbf{x}(t)$.
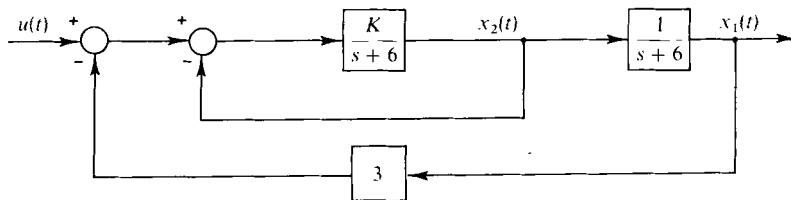
FIG. 2.P1   Blocks diagram for Problem 2.4.

2.4 A system can be represented by the block diagram of Fig. 2.P1.

(a) Determine the state equations in terms of $x_1(t)$ and $x_2(t)$.

(b) Determine the transfer function $\{x_1(s)/u(s)\}$.

(c) As a designer, you can control the gain K in the block diagram $(K > 0)$. Describe how the system dynamics are affected by letting K be changed.

(d) Determine the canonical form state space description of the system. If and where appropriate, express in modified canonical form. Describe appropriate $\mathbf{F}$, $\mathbf{b}$, and $\mathbf{h}^T$ without excessive algebra.

2.5 For a state equation of the form

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_1 & -a_2 & -a_3 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u$$

determine the resolvent matrix $\mathbf{\Phi}(s)$. Now let the output be expressed as

$$y = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \mathbf{x}$$

Determine the transfer function relating y to u. What is the differential equation that corresponds to this transfer function or state space description?

What is the steady state value of y to a unit impulse? To a unit step? Under what conditions are these evaluations valid? Show that if $c_1 = 1$ and $c_2 = c_3 = 0$, then the preceding system model is both observable and controllable for all values of a_1, a_2, and a_3.

2.6 A state space model of the system given in Fig. 2.P2 can be expressed in the form of the minimal realization

$$\frac{d}{dt}\begin{bmatrix} V(t) \\ f(t) \end{bmatrix} = \begin{bmatrix} -B/m & -1/m \\ 2K & 0 \end{bmatrix}\begin{bmatrix} V(t) \\ f(t) \end{bmatrix} + \begin{bmatrix} 1/m \\ 0 \end{bmatrix} F_s(t)$$

$$\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} V(t) \\ f(t) \end{bmatrix}$$

where V = velocity of the mass and f = force through an equivalent spring of stiffness $2K$. Suppose you want to generate another minimal realization in terms of the variables f and df/dt. How would
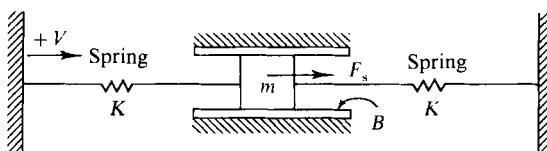

FIG. 2.P2   Diagram for system of Problem 2.6.

you do it? Write this form explicitly. Generate the similarity transformation matrix that relates these two realizations explicitly. Write the solution to these equations for df/dt, in both the time domain and the frequency domain.

2.7 Figure 2.P3 depicts a "stable platform," a system that is composed of a motor-driven gimbal, and which attempts to keep the platform rotating at a commanded angular velocity with respect to inertial space (often zero) despite interfering torques due to base motions or forces applied to the platform. Sensed angular velocity from a gyro on the platform is used by the controller to generate appropriate commands to the gimbal motor to achieve this purpose. Consider a single axis stable platform with direct drive and no tachometer feedback. Assume that its components can be described by the following relationships, where D indicates the time differentiation operation.

Controlled member: $I_{cm}D\omega_{im} = M_m + M_{intf}$.

Gyro equation: $De_{sig} = S_{sg}[\omega_{cmd} - \omega_{im} + \omega_d]$.

Control equation: $M_m = [S_c/S_{sg}]F_c(D)e_{sig}$.

The variables are defined as I_{cm}, moment of inertia of controlled member; ω_{im}, angular velocity of controlled member with respect to the inertial frame of reference; M_m, servo motor torque; M_{intf}, interfering torque; e_{sig}, signal voltage (gyro output); ω_{cmd}, commanded angular velocity; ω_d, drift rate of gyro; S_{sg}, signal generator gain; S_c, control gain; and $F_c(D)$, transfer function of control system.

(a)   Show that the general performance equation for the stable platform is

$$[I_{cm}D^2 + S_cF_c(D)]\omega_{im} = S_cF_c(D)[\omega_{cmd} + \omega_d] + DM_{intf}$$

(b)   Draw a system block diagram corresponding to the three component equations, and thereby to the general performance equation.

(c)   Using the outputs of the integrators in this diagram as state variables, write out a state space description of the system for the case of $F_c(D) = 1$.

(d)   To attain a controlled member "stable" with respect to inertial space, ω_{cmd} is set equal to zero. Using the result of part (c), evaluate the effect of a constant drift rate on the angular rate ω_{im} of the controlled member. Also evaluate the effect of a constant interfering torque on ω_{im}.

Platform or "controlled member"
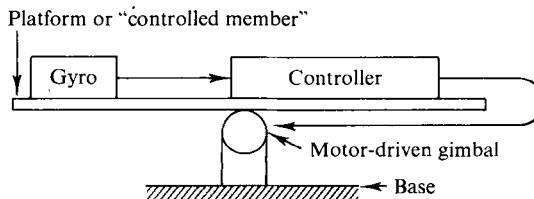
FIG. 2.P3   One-axis stable platform configuration.

2.8   This problem deals with a pitch attitude control system (PACS) for a rigid rocket vehicle employing a gimbaled rocket motor to obtain pitching torques. A pitching moment is obtained by tilting the rocket motor to give a transverse component of thrust, applied behind the center of gravity. The rocket motor is tilted in its gimbals by means of a hydraulic positioning system described by a transfer function of

$$0.1 \left/ \left[\frac{s}{250} + 1\right]\left[\frac{s^2}{(365)^2} + \frac{2(0.616)}{365}s + 1\right]\right. \quad \text{deg/V}$$

Two gyroscopes are used to produce measurements for feedback. One is an attitude gyro which produces an electrical signal proportional to pitch angle (with proportionality constant K_{ag} V/deg),

and the other is a rate gyro which develops an electrical signal proportional to pitch rate (proportionality constant K_{rg} V-sec/deg). Attenuation controls enable you to vary the pitch rate signal level relative to the pitch angle signal level. The objective of a control system design would be to find the best ratio of these signals. These two signals are added, and their sum is then subtracted from a pitch angle command voltage, all by means of a summing amplifier. The pitch angle command voltage is produced by a guidance computer. The output of the summing amplifier is used as the actuating signal for the rocket motor positioning system.

(a) Prepare a block diagram for the PACS.

(b) Generate a time-domain model for this system in the form of a state differential equation and output relation.

(c) Generate a frequency-domain model in the form of a Laplace transform transfer function.

2.9 Consider the following system model (note that it is not in standard observable or standard controllable form):

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u(t)$$

$$z(t) = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \mathbf{x}(t)$$

(a) Draw a block diagram of this state space model, explicitly labeling the states, input, and output.

(b) Is the system model completely controllable?

(c) Is the system model completely observable?

(d) Find the transfer function of the system, $G(s)$, that relates the output, $z(t)$, to the input, $u(t)$, through the observable and controllable portion of the system model.

(e) One system pole is located at $s = -1$. Plot the poles and zeros of $G(s)$. Is this a "nonminimum phase" and stable system—i.e., are there singularities in the right half plane?

(f) Develop the canonical form state equations to describe this system (using the transfer function directly is probably the most straightforward means).

(g) Can you determine a transformation matrix, T, that will transform the given state equations into the canonical form equations?

(h) Assuming the system to be initially at rest (all appropriate initial conditions zero), find the system response, $z(t)$, to a unit impulse input, $u(t)$.

(i) Under the same assumptions of initial rest, compute $z(t)$ for a unit step input, $u(t)$.

There are a number of valid methods of attaining the answers to these questions. One matrix form that can be useful in some parts (though not necessarily required in any part) would be adj$(f\mathbf{I} - \mathbf{F})$, where f is some appropriate quantity and $\mathbf{F}$ is the given $\mathbf{F}$ matrix in the problem; the evaluation of this form is

$$\text{adj}(f\mathbf{I} - \mathbf{F}) = \begin{bmatrix} f^2 + 6f + 11 & f + 6 & 1 \\ -6 & f^2 + 6f & f \\ -6f & -11f - 6 & f^2 \end{bmatrix}$$

2.10 Explain why the following system realizations are or are not both observable and controllable.

(a)

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & -5 & -4 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u}(t)$$

$$z(t) = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \mathbf{x}(t)$$

(b)

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} -3 & 0 \\ 0 & -5 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \mathbf{u}(t)$$

$$\mathbf{z}(t) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \mathbf{x}(t)$$

(c)

$$\dot{\mathbf{x}}(y) = \begin{bmatrix} -3 & 0 \\ 0 & -5 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \mathbf{u}(t)$$

$$\mathbf{z}(t) = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \mathbf{x}(t)$$

2.11   Calculate the controllability Gramian $\mathbf{W}(0, T)$ for the driven oscillator:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

Is this system model completely controllable? Compare this to the answer achieved using $\mathbf{W}_{\mathrm{Tl}}$.

2.12   Consider the single-axis error model for an inertial navigation system (INS) shown in Fig. 2.P4. In this simplified model, gyro drift rate errors are ignored. It is desired to estimate the vertical deflection process, e_ξ. Consider the system state vector consisting of the four states:

$$\mathbf{x}^T = \begin{bmatrix} e_{\mathrm{pb}} & \delta_\mathrm{p} & \delta_\mathrm{v} & e_{\xi b} \end{bmatrix}$$

(a)   For position measurements only (z_v not available), set up the observability matrix and determine whether this system model is observable.

(b)   Now assume we have both position and velocity measurements. Is this system model observable?

(c)   Now assume $e_{\mathrm{pb}} = 0$ and can be eliminated from the state error model. Is this system model observable with position measurements only?
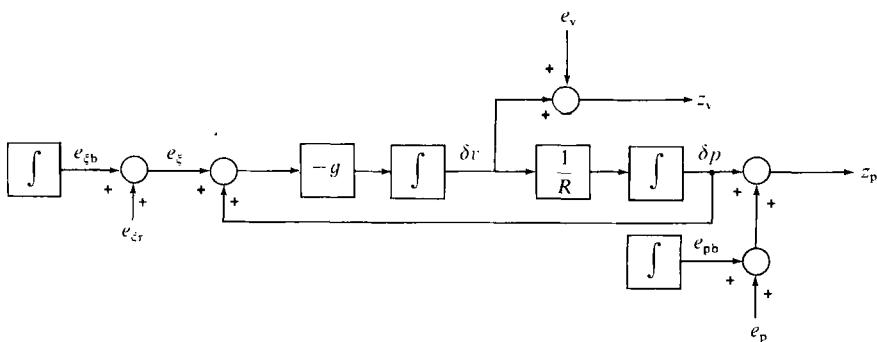

FIG. 2.P4   Single-axis error model for INS. $e_{\xi r}$ = unbiased random vertical deflection, $e_{\xi b}$ = vertical deflection bias, e_v = uncorrelated velocity measurement error, e_{pb} = bias position measurement error, e_p = uncorrelated position measurement error. From *Applied Optimal Estimation* by A. Gelb (ed.). © 1974. Used with permission of the M.I.T. Press.

2.13 Show that the state transition matrix corresponding to the **F** in Example 2.8 for a circular nominal trajectory with $r_0 = 1$ is given by:

$$\Phi(t, 0) = \begin{bmatrix} 4 - 3\cos\omega t & (\sin\omega t)/\omega & 0 & 2(1 - \cos\omega t)/\omega \\ 3\omega\sin\omega t & \cos\omega t & 0 & 2\sin\omega t \\ 6(-\omega t + \sin\omega t) & -2(1 - \cos\omega t)/\omega & 1 & (-3\omega t + 4\sin\omega t)/\omega \\ 6\omega(-1 + \cos\omega t) & -2\sin\omega t & 0 & -3 + 4\cos\omega t \end{bmatrix}$$

2.14 Given that **F** is a 2×2 constant matrix and given that

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t)$$

Suppose that if

$$\mathbf{x}(0) = \begin{bmatrix} 1 \\ -3 \end{bmatrix} \quad \text{then} \quad \mathbf{x}(t) = \begin{bmatrix} e^{-3t} \\ -3e^{-3t} \end{bmatrix}$$

and if

$$\mathbf{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{then} \quad \mathbf{x}(t) = \begin{bmatrix} e^t \\ e^t \end{bmatrix}$$

Determine the transition matrix for the system and the matrix **F**.

Problem 2.14 is from *Finite Dimensional Linear Systems* by R. Brockett. © 1970. Used with permission of John Wiley & Sons, Inc.

2.15 (a) Show that, for all t_0, t_1, and t,

$$\Phi(t, t_0) = \Phi(t, t_1)\Phi(t_1, t_0)$$

by showing that both quantities satisfy the same linear differential equation and "initial condition" at time t_1. Thus, the solution to $\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t)$ with $\mathbf{x}(t_0) = \mathbf{x}_0$ (i.e., $\Phi(t, t_0)\mathbf{x}_0$) at any time t_2 can be obtained by forming $\mathbf{x}(t_1) = \Phi(t_1, t_0)\mathbf{x}_0$ and using it to generate $\mathbf{x}(t_2) = \Phi(t_2, t_1)\mathbf{x}(t_1)$.

(b) Since it can be shown that $\Phi(t, t_0)$ is nonsingular, show that the above "semigroup property" implies that

$$\Phi^{-1}(t, t_0) = \Phi(t_0, t)$$

2.16 Let $\Phi(t, t_0)$ be the state transition matrix associated with $\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)$. Consider a change of variables,

$$\mathbf{x}^*(t) = \mathbf{T}(t)\mathbf{x}(t)$$

where $\mathbf{T}(\cdot)$ is differentiable and $\mathbf{T}^{-1}(t)$ exists for all time t. Show that the state transition matrix associated with the transformed state variables, $\Phi^*(t, t_0)$, is the solution to

$$\dot{\Phi}^*(t, t_0) = [\mathbf{T}(t)\mathbf{F}(t)\mathbf{T}^{-1}(t) + \dot{\mathbf{T}}(t)\mathbf{T}^{-1}(t)]\Phi^*(t, t_0)$$
$$\Phi^*(t_0, t_0) = \mathbf{I}$$

and that

$$\Phi(t, t_0) = \mathbf{T}^{-1}(t)\Phi^*(t, t_0)\mathbf{T}(t_0)$$

Note that if **T** is constant, this yields a similarity transformation useful for evaluation of the state transition matrix.

2.17 Let $\mathbf{F}$ be constant. Then the evaluation of $\boldsymbol{\Phi}(t, t_0) = \boldsymbol{\Phi}(t - t_0)$ can be obtained by
(a) approximation through truncation of series definition of matrix exponential, $e^{\mathbf{F}(t - t_0)}$:

$$e^{\mathbf{F}(t - t_0)} = \mathbf{I} + \mathbf{F}(t - t_0) + \frac{1}{2!} \mathbf{F}^2(t - t_0)^2 + \cdots$$

(b) Laplace methods of solving $\dot{\boldsymbol{\Phi}}(t - t_0) = \mathbf{F}\boldsymbol{\Phi}(t - t_0)$, $\boldsymbol{\Phi}(0) = \mathbf{I}$:

$$\boldsymbol{\Phi}(t - t_0) = \mathscr{L}^{-1}\{[s\mathbf{I} - \mathbf{F}]^{-1}\}|_{(t - t_0)}$$

where $\mathscr{L}^{-1}\{\cdot\}|_{(t-t_0)}$ denotes inverse Laplace transform evaluated with time argument equal to $(t - t_0)$.

(c) Cayley–Hamilton theorem (for $\mathbf{F}$ with nonrepeated eigenvalues)

$$\boldsymbol{\Phi}(t - t_0) = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{F} + \alpha_2 \mathbf{F}^2 + \cdots + \alpha_{n-1} \mathbf{F}^{n-1}$$

To solve for the n functions of $(t - t_0)$, $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, the n eigenvalues of $\mathbf{F}$ are determined as $\lambda_1, \ldots, \lambda_n$. Then

$$e^{\lambda(t - t_0)} = \alpha_0 + \alpha_1 \lambda + \alpha_2 \lambda^2 + \cdots + \alpha_{n-1} \lambda^{n-1}$$

must be satisfied by each of the eigenvalues, yielding n equations for the n unknown α_i's.

(d) Sylvester expansion theorem (for $\mathbf{F}$ with nonrepeated eigenvalues)

$$\boldsymbol{\Phi}(t - t_0) = \mathbf{F}_1 e^{\lambda_1(t - t_0)} + \mathbf{F}_2 e^{\lambda_2(t - t_0)} + \cdots + \mathbf{F}_n e^{\lambda_n(t - t_0)}$$

where λ_i is the ith eigenvalue of $\mathbf{F}$ and $\mathbf{F}_i$ is given as the following product of $(n - 1)$ factors:

$$\mathbf{F}_i = \left[\frac{\mathbf{F} - \lambda_1 \mathbf{I}}{\lambda_i - \lambda_1}\right] \cdots \left[\frac{\mathbf{F} - \lambda_{i-1}\mathbf{I}}{\lambda_i - \lambda_{i-1}}\right]\left[\frac{\mathbf{F} - \lambda_{i+1}\mathbf{I}}{\lambda_i - \lambda_{i+1}}\right] \cdots \left[\frac{\mathbf{F} - \lambda_n \mathbf{I}}{\lambda_i - \lambda_n}\right]$$

Use the four methods to evaluate $\boldsymbol{\Phi}(t - t_0)$ if $\mathbf{F}$ is given by

$$\mathbf{F} = \begin{bmatrix} 0 & 6 \\ -1 & -5 \end{bmatrix}$$

Let $(t - t_0) = 0.1$ sec; if the series method of part (a) is truncated at the first order term, what is the greatest percentage error (i.e., [(calculated value $-$ true)/true value] $\cdot$ 100%) committed in approximating the four elements of $\boldsymbol{\Phi}(0.1)$? What if it were truncated at the second order term?

2.18 Given a homogeneous linear differential equation $\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t)$, the associated "adjoint" differential equation is the differential equation for the n-vector $\mathbf{p}(t)$ such that the inner product of $\mathbf{p}(t)$ with $\mathbf{x}(t)$ is constant for all time:

$$\mathbf{x}(t)^{\mathrm{T}}\mathbf{p}(t) = \text{const}$$

(a) Take the derivative of this expression to show that the adjoint equation associated with $\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t)$ is

$$\dot{\mathbf{p}}(t) = -\mathbf{F}^{\mathrm{T}}(t)\mathbf{p}(t)$$

(b) If $\boldsymbol{\Phi}_x(t, t_0)$ is the state transition matrix associated with $\mathbf{F}(t)$ and $\boldsymbol{\Phi}_p(t, t_0)$ is the state transition matrix associated with $[-\mathbf{F}^{\mathrm{T}}(t)]$, then show that

$$\boldsymbol{\Phi}_p(t, t_0) = \boldsymbol{\Phi}_x^{\mathrm{T}}(t_0, t) = [\boldsymbol{\Phi}_x^{\mathrm{T}}(t, t_0)]^{-1}$$

To do this, show that $[\boldsymbol{\Phi}_p^{\mathrm{T}}(t, t_0)\boldsymbol{\Phi}_x(t, t_0)]$ and $\mathbf{I}$ satisfy the same differential equation and initial condition.

(c)   Show that, as a function of its second argument, $\mathbf{\Phi}_x(t, \tau)$ must satisfy

$$\partial[\mathbf{\Phi}_x(t, \tau)]/\partial\tau = -\mathbf{\Phi}_x(t, \tau)\mathbf{F}(\tau)$$

or, in other words,

$$\partial[\mathbf{\Phi}_x{}^{\mathrm{T}}(t, \tau)]/\partial\tau = [-\mathbf{F}(\tau)^{\mathrm{T}}]\mathbf{\Phi}_x{}^{\mathrm{T}}(t, \tau)$$

2.19   The Euler equations for the angular velocities of a rigid body are

$$I_1\dot{\omega}_1 = (I_2 - I_3)\omega_2\omega_3 + u_1$$
$$I_2\dot{\omega}_2 = (I_3 - I_1)\omega_1\omega_3 + u_2$$
$$I_3\dot{\omega}_3 = (I_1 - I_2)\omega_1\omega_2 + u_3$$

Here ω is the angular velocity in a body-fixed coordinate system coinciding with the principal axes, $\mathbf{u}$ is the applied torque, and I_1, I_2, and I_3 are the principal moments of inertia. If $I_1 = I_2$, we call the body symmetrical. In this case, linearize these equations about the solution $\mathbf{u} = \mathbf{0}$,

$$\begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \\ \dot{\omega}_3 \end{bmatrix} = \begin{bmatrix} \cos \omega_0 (I_2 - I_3)/I_1 \\ \sin \omega_0 (I_2 - I_3)/I_1 \\ \omega_0 \end{bmatrix}$$

Problem 2.19 is from *Finite Dimensional Linear Systems* by R. Brockett.   c   1970. Used with permission of John Wiley & Sons, Inc.

2.20   Consider a system with input $r(t)$ and output $c(t)$ modeled by the nonlinear differential equation

$$\ddot{c}(t) + c^3(t)\dot{c}^2(t) + \sin[c(t)] - t^2 c(t) = r(t)$$

Determine the linear perturbation equations describing the system's behavior near nominal trajectories of

(1) $c(t) = r(t) = 0$,

(2) $c(t) = t, r(t) = \sin t$.

Describe how you would obtain the equivalent discrete-time model for the perturbation equations in the two preceding cases, to describe the perturbed output Δc at discrete time points:

$$\Delta c(kT), \qquad k = 0, 1, 2, \ldots, \qquad \text{for} \quad \text{given } T$$

2.21   Consider the system configuration of Fig. 2.P5. The sampler has a period of T seconds and generates the sequence $\{e_1(0), e_1(T), e_1(2T), \ldots\}$. The algorithm implemented in the digital computer is a first-order difference approximation to differentiation:

$$e_2(t_i) = [e_1(t_i) - e_1(t_{i-1})]/T$$

Finally, the zero-order hold (ZOH) generates a piecewise constant output by holding the value of its input over the ensuing sample period:

$$u(t) = e_2(t_i) \qquad \text{for all} \quad t \in [t_i, t_{i+1})$$

(a)   Generate the equivalent discrete-time model for the "plant."

(b)   Develop the discrete-time state equation and output relation model of the entire system configuration, and thus describe $c(t_i) = c(iT)$ for $i = 0, 1, 2, \ldots$.

(c)   Determine the eigenvalues of the overall system state transition matrix; if the magnitude of any of these is greater than one, the system is unstable.
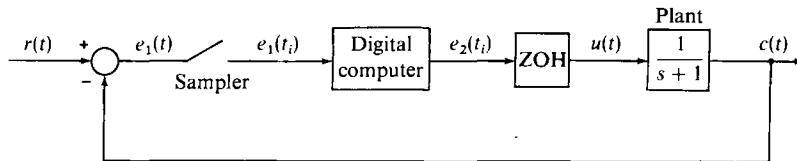
FIG. 2.P5   Block diagram for Problem 2.21.
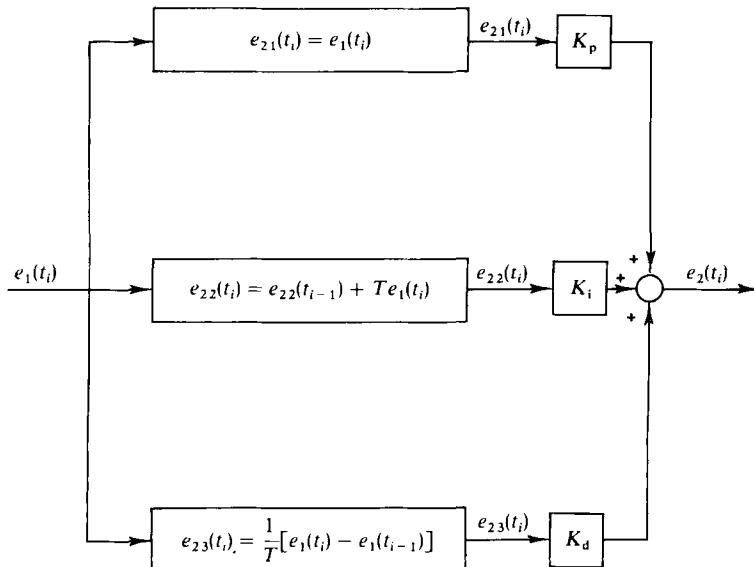
FIG. 2.P6   Digital PID controller.

2.22   A PID (proportional plus integral plus derivative) controller is a device which accepts a signal $e_1(t)$ as input and generates an output as

$$e_2(t) = K_p e_1(t) + K_i \int_{t_0}^{t} e_1(\tau)\,d\tau + K_d \dot{e}_1(t)$$

where the coefficients K_p, K_i, and K_d are adjusted to obtain desirable behavior from the closed loop system generated by using a PID controller for feedback.

A simple digital PID controller for use with an iteration period of T seconds is shown in Fig. 2.P6. Show that the difference equation approximation to the integration operation is the result of Euler integration, and that this channel can be described by the discrete-time state and output model:

$$x_i(t_{i+1}) = x_i(t_i) + e_1(t_i), \qquad e_{22}(t_i) = Tx_i(t_i) + Te_1(t_i)$$

Similarly show that the "derivative" first order difference approximation can be represented as

$$x_d(t_{i+1}) = e_1(t_i), \qquad e_{23}(t_i) = -[x_d(t_i)/T] + e_1(t_i)/T$$

Generate the state vector difference equation and output relation model for this digital PID controller. In general, the relationships among scalar difference equations, discrete-time state equations and output relations, and Z-transform transfer functions (not discussed in this book) are analogous to the relationships among scalar differential equations, differential state models, and Laplace transform transfer functions (where applicable), and methods analogous to those of this chapter yield means of generating one form from the other.
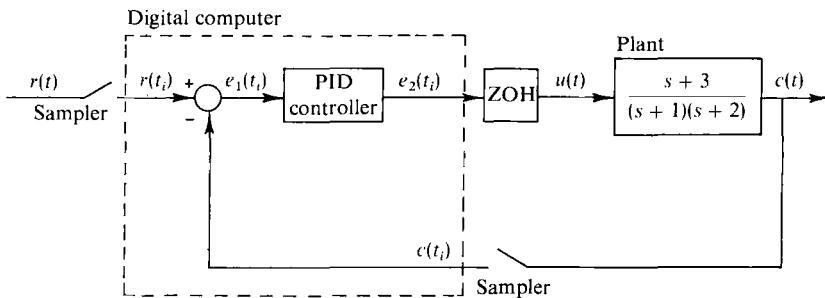
FIG. 2.P7   Control system configuration for Problem 2.23.

2.23   Consider the application of the PID controller of the previous problem applied to the system configuration depicted in Fig. 2.P7. Note that the samplers are synchronized and have sample period T. The zero-order hold (ZOH) generates

$$u(t) = e_2(t_i) \qquad \text{for all} \quad t \in [t_i, t_{i+1})$$

(a)   Develop the equivalent discrete-time model for the "plant."

(b)   Develop the discrete-time state model of the entire control system configuration, and thereby means of generating

$$c(iT) \qquad \text{for} \quad i = 0, 1, 2, \ldots$$

(c)   To evaluate adequacy of controller performance, one desires knowledge of $c(t)$ for all time between sample times as well. Generate the relations necessary to evaluate $c(t)$ for all $t \in [iT, (i+1)T]$.

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

59

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 $\Omega: \omega \in \Omega$. 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 $\Omega: A \subset \Omega$. 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. ■
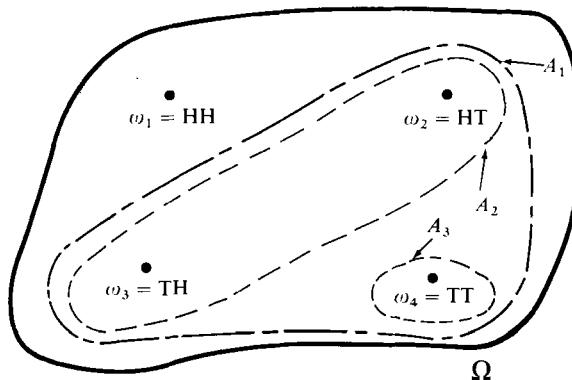
FIG. 3.1   Two tosses of a coin.
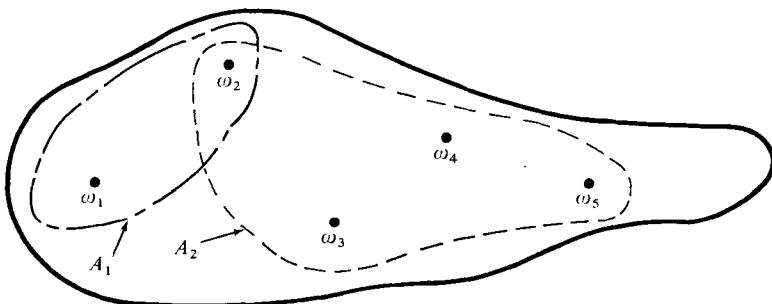
The sample space Ω can be discrete, with a finite or countably infinite number of elements, such as the space for the coin toss experiment described in the previous example. On the other hand, it could also be continuous, with an uncountable number of elements, such as the space appropriate to describe 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 $\mathscr{F}$. In other words, the sets A_1, A_2, $A_3, \ldots$, which will be admissible for consideration will be elements of the class $\mathscr{F}: A_i \in \mathscr{F}$.

A σ-algebra $\mathscr{F}$ 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 $\mathscr{F}$ (i.e., if $A_i \in \mathscr{F}$), then:

(1) $A_i^* \in \mathscr{F}$, where A_i^* is the complement of A_i, $A_i^* = \Omega - A_i$,
(2) $\Omega \in \mathscr{F}$ [and then the empty set $\varnothing \in \mathscr{F}$ also, due to the preceding (1)],

FIG. 3.2   Sample space Ω.

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

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

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

Whereas Ω is a collection of points (elementary outcomes, ω), $\mathscr{F}$ 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 $\mathscr{F}$ be the class of sets generated by sets of the form (each of which is a subset of Ω):

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

and their complements, unions, and intersections. The notation in Eq. (3-2) requires some explanation. In other words, A is the set of ω'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 (:) $\omega \leq \mathbf{a}$, where ω and $\mathbf{a}$ are n-dimensional vectors and $\mathbf{a}$ is specified. Furthermore, $\omega \leq \mathbf{a}$ is to be interpreted componentwise: $\omega \leq \mathbf{a}$ 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 ω 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 $\mathscr{F}_B$. 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 $\mathscr{F}_B$ 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

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

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 $\mathscr{F}$ that assigns a value, $P(A)$, to each A which is a member of $\mathscr{F}$ ($A \in \mathscr{F}$) such that:

(1) $P(A) \ge 0$ for all $A \in \mathscr{F}$,

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

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

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

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 $A \in \mathscr{F}$ does correspond to one's intuition of probability gained through the concept of relative frequency of occurrence. Each set of interest (i.e., each $A \in \mathscr{F}$) is assigned a probability value between 0 and 1 (P is a mapping from $\mathscr{F}$ 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,
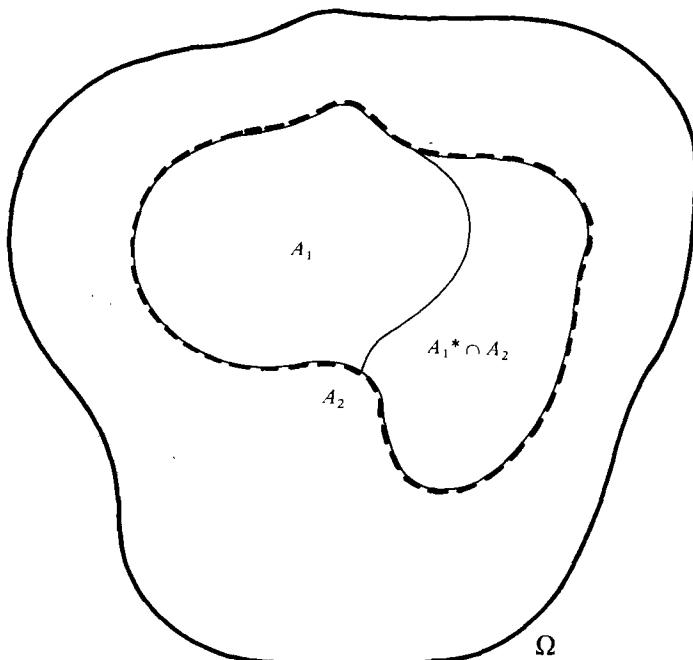
FIG. 3.3 $A_1 \subset A_2$ implies $P(A_1) \leq P(A_2)$.

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 $(\Omega, \mathscr{F}, P)$ 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 $\mathscr{F}_1$ be composed of $\varnothing$, $\Omega = \{1, 2, 3, 4, 5,$ 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 $\mathscr{F}_1$. 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 $\mathscr{F}_2$ be composed of $\varnothing$, Ω, A_1, A_2, and all possible complements, unions, and intersections thereof. Thus, $\mathscr{F}_2$ 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 $P(A_1) = P(\{1$ 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 $\mathscr{F}$ 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, $\mathscr{F}$ 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

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

for ξ any value on the real line ($\xi \in R^1$), is an element of the σ-algebra $\mathscr{F}$ (i.e., $A \in \mathscr{F}$). 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 $\{\omega : x(\omega) \le \xi\}$ is meant to read, "the set of ω in Ω such that the values assumed by the random variable function $x(\cdot)$, for those ω as its argument, $x(\omega) = x$, 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

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

for any $\boldsymbol{\xi} \in R^n$, is an element of $\mathscr{F}$. 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 $\mathscr{F}$. 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 $\{\mathbf{x}(\omega) \in R^n : -\infty < x_i(\omega) \le \xi_i ; i = 1, 2, \ldots, n\}$ 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 $\mathscr{F}$ is made up of all possible complements, unions, and intersections of these two intervals, so that

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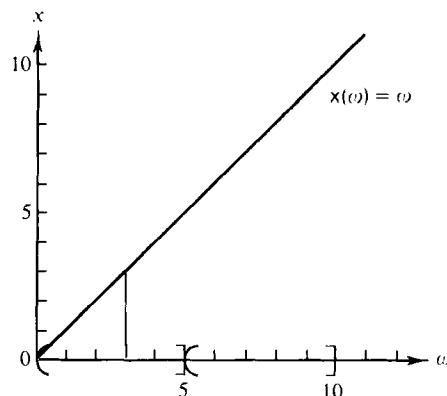

FIG. 3.4   Function that is not a random variable.

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 $x(\cdot)$ such that $x(\omega) = \omega$, as in Fig. 3.4. This is not a suitable choice. Choose, for example, $\xi = 3$, as shown. Then the set A defined by

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

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

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

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,

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

are all elements of $\mathscr{F}$. ∎
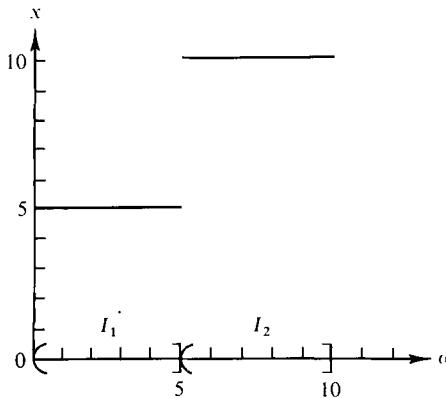
FIG. 3.5 Random variable definition. $x(\omega) = 5$ if $\omega \in I_1$ and $x(\omega) = 10$ if $\omega \in I_2$.

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

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

are events in Ω $(A_i \subset \Omega)$ for which probabilities have been defined $(A_i \in \mathscr{F})$.
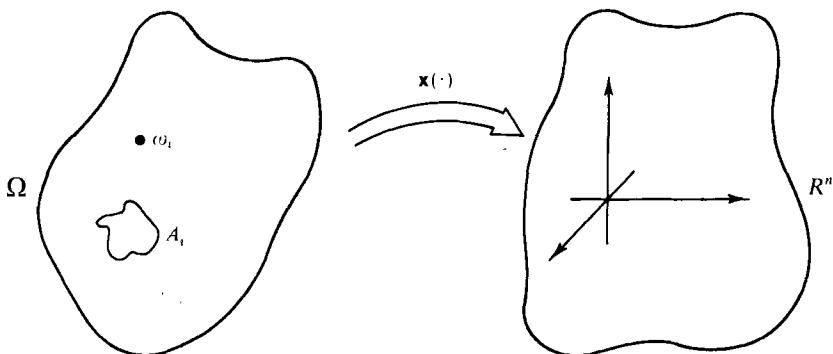
FIG. 3.6   The random variable mapping.

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 $\mathscr{F}_B$ generated by sets of the form $A_i = \{\omega : \omega \le \mathbf{a}, \omega \in \Omega\}$. 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

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

have probabilities: probabilities are defined for them because $A \subset \Omega$ and $A \in \mathscr{F}$, 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

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

$$= \text{``}P(\mathbf{x} \le \boldsymbol{\xi})\text{''} \tag{3-6b}$$

$$= \text{``}P(\mathbf{x}_1 \le \xi_1, \mathbf{x}_2 \le \xi_2, \dots, \mathbf{x}_n \le \xi_n)\text{''} \tag{3-6c}$$

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

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

this is sometimes called the joint probability distribution function of $\mathbf{x}_1, \mathbf{x}_2, \dots,$ and $\mathbf{x}_n$.
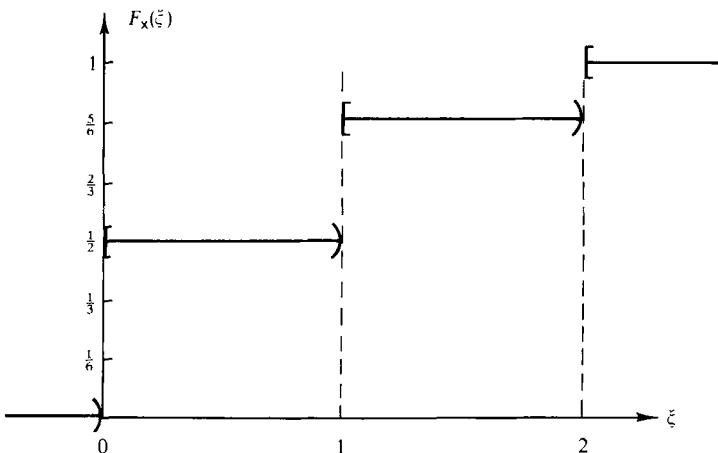
FIG. 3.7 Probability distribution function.

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

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

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

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

For $\xi < 0$, $P(\{\omega : x(\omega) \leq \xi\}) = P(\varnothing) = 0$.
For $0 \leq \xi < 1$, $P(\{\omega : x(\omega) \leq \xi\}) = P(\{A_1 \cup A_2\}^*) = 1 - P_1 - P_2 = \frac{1}{2}$.
For $1 \leq \xi < 2$, $P(\{\omega : x(\omega) \leq \xi\}) = P(A_2^*) = 1 - P_2 = \frac{5}{6}$.
For $2 \leq \xi < \infty$, $P(\{\omega : x(\omega) \leq \xi\}) = P(\Omega) = 1$.

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 $\mathscr{F}$ ($A \in \mathscr{F}$) called a σ-algebra. For any $A \in \mathscr{F}$, we could evaluate the *set function* $P(\cdot)$, a mapping from $\mathscr{F}$ into $[0, 1]$, to generate probabilities as $P(A)$. The triplet $(\Omega, \mathscr{F}, P)$ then defined what was termed a probability space.

We also defined a *point function* $x(\cdot)$ called a random variable, a mapping from Ω into R^n, which could be evaluated for each $\omega \in \Omega$ to yield realizations $x(\omega)$. The probabilities established as $P(A)$ and the realizations $x(\omega)$ of the random variable x are then related by the probability distribution function $F_x(\cdot)$, a mapping from R^n into $[0, 1]$, that yields $F_x(\xi)$ as the probability of the set of $\omega \in \Omega$ such that $x(\omega) \leq \xi$.
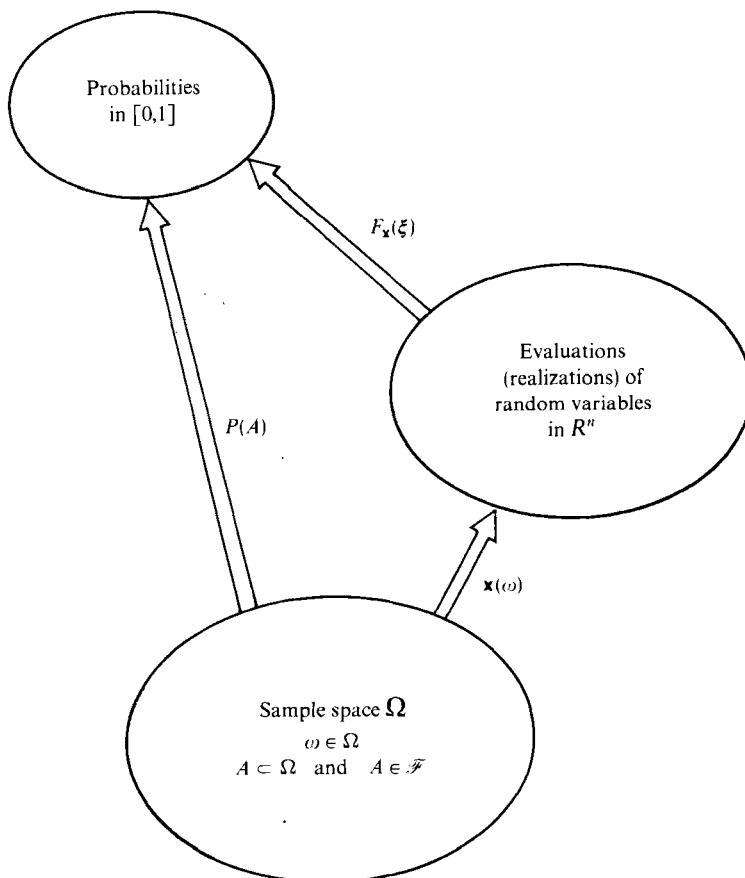
FIG. 3.8   Probability and random variables.

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 **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 $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_n]^{\mathrm{T}}$:

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

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

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 $\mathbf{x}_i(\omega) \le 2$ must be at least as large as the probability of the set of ω such that $\mathbf{x}_i(\omega) \le 1$.

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

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

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

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:

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

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 $\{\omega : \mathbf{x}(\omega) \le \xi_2\}$ can be decomposed into the union of two disjoint sets:

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

Because the sets on the right are disjoint, we have

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

so we can write

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

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 $x(\cdot)$ 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

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

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,

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

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 $f_x(\cdot)$ exists such that

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

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

$$F_x(\boldsymbol{\xi}) = \int_{-\infty}^{\boldsymbol{\xi}} f_x(\boldsymbol{\rho}) \, d\boldsymbol{\rho} \tag{3-16b}$$

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

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

$$f_x(\boldsymbol{\xi}) = \frac{\partial^n}{\partial \xi_1 \, \partial \xi_2 \, \cdots \, \partial \xi_n} F_x(\boldsymbol{\xi}) \tag{3-17}$$

This relationship and (3-16), combined with properties of F_x, yield some properties of f_x. Since F_x is monotonic nondecreasing,

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

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

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

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

$$f_{x_1, \ldots, x_k}(\xi_1, \ldots, \xi_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{x_1, \ldots, x_n}(\xi_1, \ldots, \xi_k, \ldots, \xi_n)\, d\xi_{k+1} \cdots d\xi_n$$

$$(3\text{-}20)$$

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

$$P(\{\omega : \mathbf{x}(\omega) = \xi_0\}) = 0 \qquad (3\text{-}21)$$

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

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

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 $\mathbf{x}(\omega)$ lies in the infinitesimal interval from ξ_1 to $(\xi_1 + d\xi_1)$ is just

$$P(\{\omega : \mathbf{x}(\omega) \in [\xi_1, \xi_1 + d\xi_1]\}) = f_x(\xi_1)\, d\xi_1 \qquad (3\text{-}23a)$$

This generalizes to the n-dimensional case as

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

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

(a)

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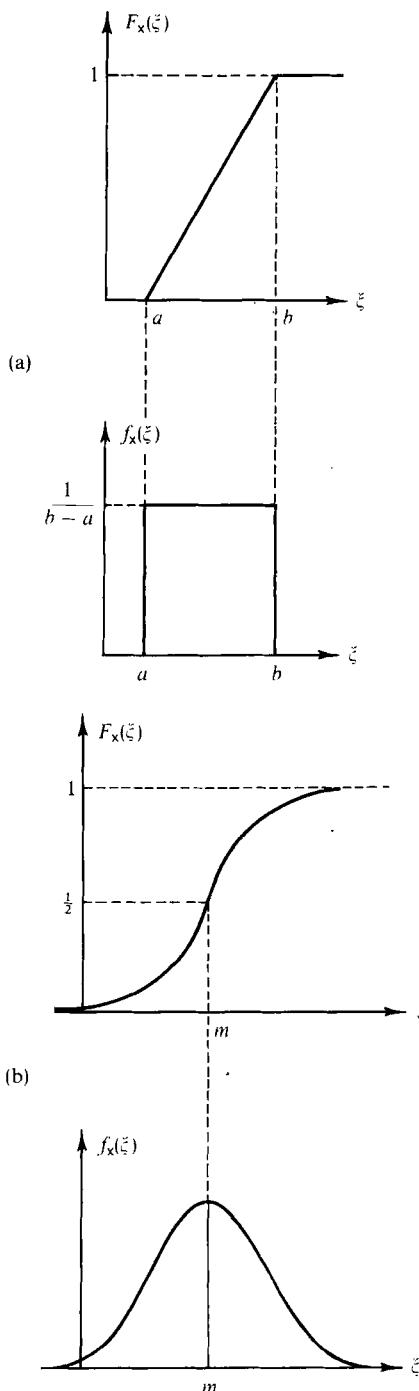
FIG. 3.9   (a) Uniform and (b) Gaussian (normal) distributions and densities.

the set A is

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

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

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), $\mathbf{x}$ is scalar and $f_\mathbf{x}(\xi)$ is a simple curve plotted against ξ. Sets A of interest are intervals along the abscissa (possibly disjoint), and $P(\{\omega:\mathbf{x}(\omega) \in A\})$ 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
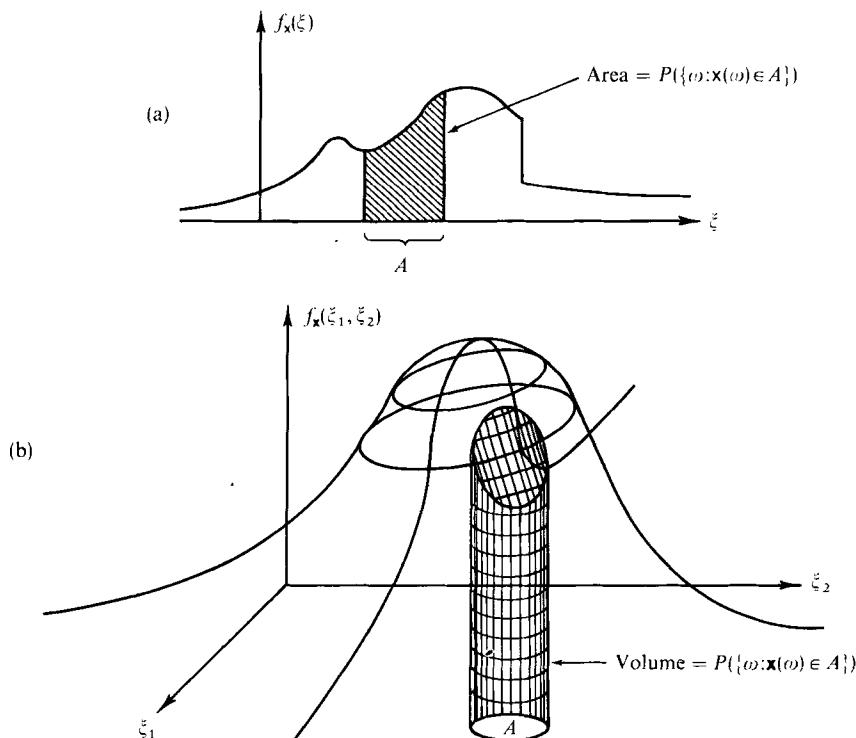
FIG. 3.10   Probability of sets. (a) Scalar-valued x. (b) Two-dimensional vector-valued **x**.

defined everywhere), the preceding ideas can be extended through

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

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

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

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

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

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:

$$(\Omega, \mathscr{F}, P) \overset{\mathbf{x}(\cdot)}{\to} (R^n, \mathscr{F}_B, P_x) \qquad (3\text{-}28)$$

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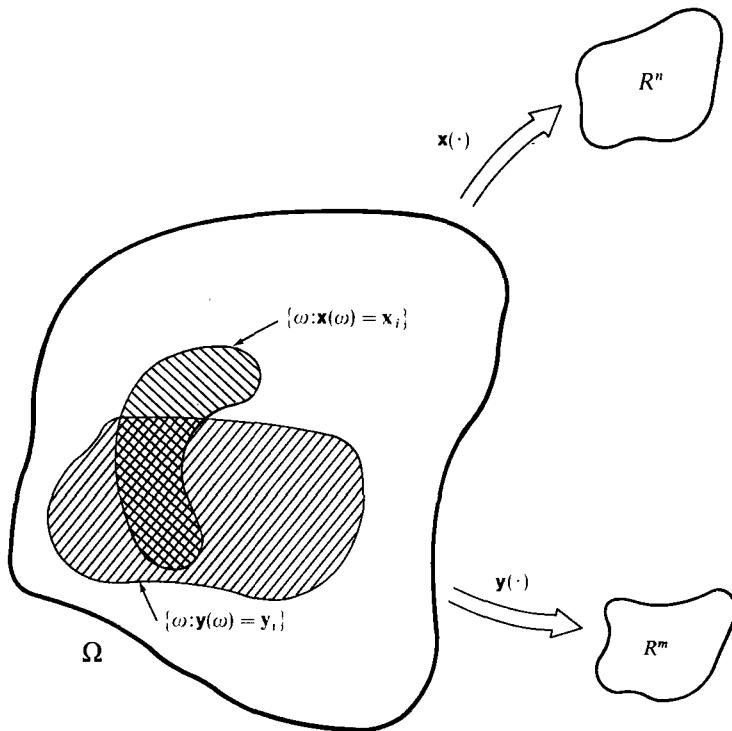
FIG. 3.11 Conditional probability via sets in Ω. The crosshatching is $\{\omega:\mathbf{x}(\omega) = \mathbf{x}_i$ and $\mathbf{y}(\omega) = \mathbf{y}_j\}$.

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 $A \in \mathscr{F}$. 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 $\{\omega:\mathbf{x}(\omega) = \mathbf{x}_i$ 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

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

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

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

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

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 $(\Omega, \mathcal{F}, P)$, or with the functions P'_{xy} and P_y' associated with $(R^{nm}, \mathcal{F}_B, P'_{xy})$ and $(R^m, \mathcal{F}_B, P_y')$ as described in (3-27). Thus, letting $f_{\mathbf{x},\mathbf{y}}(\cdot, \cdot)$ denote the joint probability density of $\mathbf{x}$ and $\mathbf{y}$,

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

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

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

$$f_{\mathbf{x}}(\xi \,|\, \mathbf{y}(\omega) \in B) = \frac{\int_B f_{\mathbf{x},\mathbf{y}}(\xi, \gamma)\, d\gamma}{\int_B f_{\mathbf{y}}(\rho)\, d\rho} \qquad (3\text{-}32)$$
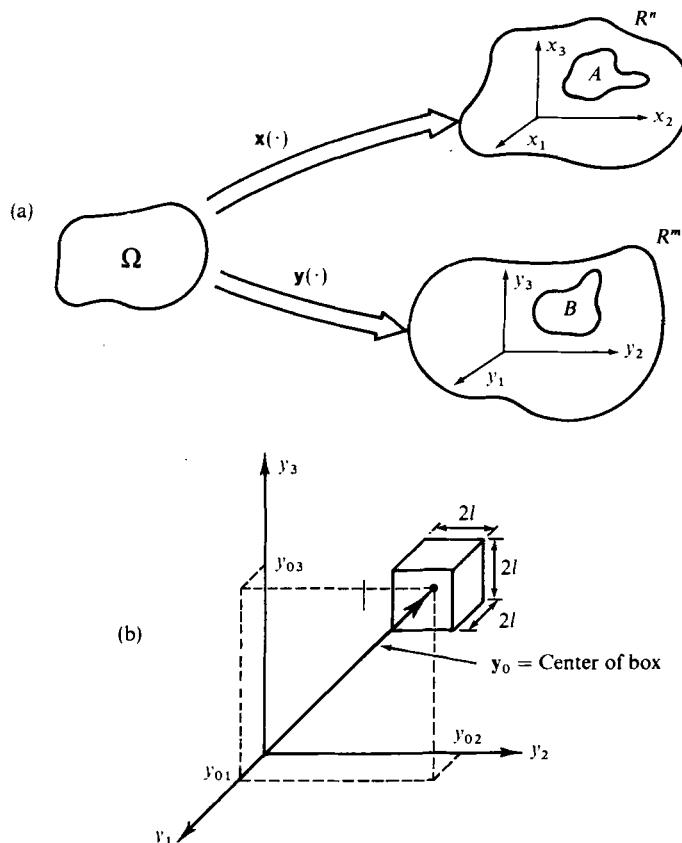
FIG. 3.12 (a) Sets of interest in R^n and R^m. (b) A particular set $B \subset R^m$. y_0 is center of box.

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:

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

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

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

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

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

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

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

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

$$\int_B \delta f_y\, d\rho = V_B \delta f_y(\mathbf{b}_2) \qquad (3\text{-}37b)$$

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

$$f_x(\xi \,|\, \mathbf{y}(\omega) \in B) = \frac{f_{x,y}(\xi, \mathbf{y}_0) + \delta f_{x,y}(\xi, \mathbf{b}_1)}{f_y(\mathbf{y}_0) + \delta f_y(\mathbf{b}_2)} \qquad (3\text{-}38)$$

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

$$\lim_{l \to 0} f_x(\xi \,|\, \mathbf{y}(\omega) \in B) = f_{x,y}(\xi, \mathbf{y}_0)/f_y(\mathbf{y}_0) \qquad (3\text{-}39)$$

defines what is meant by $f_{x|y}(\xi \,|\, \mathbf{y}_0)$, sometimes denoted as $f_x(\xi \,|\, \mathbf{y} = \mathbf{y}_0)$. 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:

$$f_{x|y}(\xi \,|\, \rho) = \frac{f_{x,y}(\xi, \rho)}{f_y(\rho)} \qquad (3\text{-}40)$$

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:

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

$$= \frac{f_y(\rho)}{f_y(\rho)} = 1 \qquad (3\text{-}41)$$

A graphical representation of (3-40) is useful for insight. Figure 3.13 portrays the joint density function $f_{x,y}(\xi, \rho)$ as a surface above the ξ–ρ plane. To generate $f_{x|y}(\xi \,|\, y_0)$ 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 $f_{x,y}(\xi, y_0)$. If that
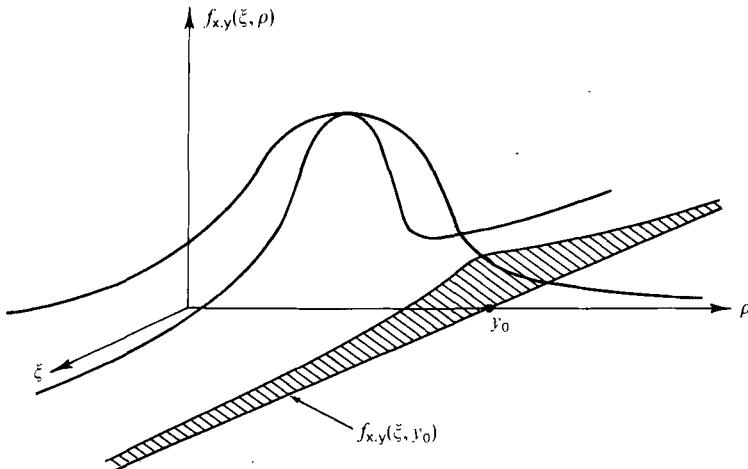
FIG. 3.13   Generation of conditional density.

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 x based upon a set of measurements $z_1 = z_1, z_2 = z_2, \ldots, z_N = z_N$. To accomplish this objective, we will propagate the conditional density $f_{x|z_1, z_2, \ldots, z_N}(\xi | z_1, z_2, \ldots, z_N)$ of the state 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

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

Through this expression, one can readily generate the conditional density $f_{x|y}(\xi | \rho)$ if it is possible to write $f_{y|x}(\rho | \xi)$ and the unconditional densities for x and y. This will be exploited to derive $f_{x|z_1, z_2, \ldots, z_N}(\xi | z_1, z_2, \ldots, z_N)$ 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:

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

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 inter-relationships among random variables. The two extremes of such relationships are independence and functional dependence, to be discussed next.

Consider two random variables, **x** mapping Ω into R^n and **y** mapping Ω into R^m, and two admissible events $A \subset R^n$ and $B \subset R^m$. Then **x** and **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:

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

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

Then, by the definition of the appropriate distribution functions,

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

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

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

for this particular choice of A and B. Thus, if **x** and **y** are independent, then

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

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

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

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

Another, more restrictive, approach is to define random vectors **x** and **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 **x** and **y** are independent, then (3-48) and Bayes' rule together yield $f_{\mathbf{x}|\mathbf{y}}(\xi|\rho)$ as

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

i.e., the conditional density for **x**, conditioned on knowledge that **y** has assumed a realization ρ, is equal to the unconditional density for **x**. This makes sense conceptually: if **x** and **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 **x** is a deterministic function of **y**, $\mathbf{x} = \phi(\mathbf{y})$, then the conditional probability density function $f_{\mathbf{x}|\mathbf{y}}(\xi|\rho)$ is an impulse:

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

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

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

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 $\xi = \phi(\rho)$. If $\mathbf{x} = \phi(\mathbf{y})$ and $\mathbf{y}(\omega) = \rho$, then $\mathbf{x}(\omega) = \phi[\mathbf{y}(\omega)] = \phi(\rho)$ with no uncertainty: all of the probability is concentrated at $\xi = \phi(\rho)$, rather than being spread over a range of ξ 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 ξ 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 $\theta(\cdot)$ 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, $\theta(\cdot)$ 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 $\theta[\mathbf{x}(\cdot)]$. Then $\mathbf{y}$ is itself a random variable, denoted as $\mathbf{y}$:

$$\mathbf{y}(\cdot) = \theta[\mathbf{x}(\cdot)] \tag{3-52a}$$
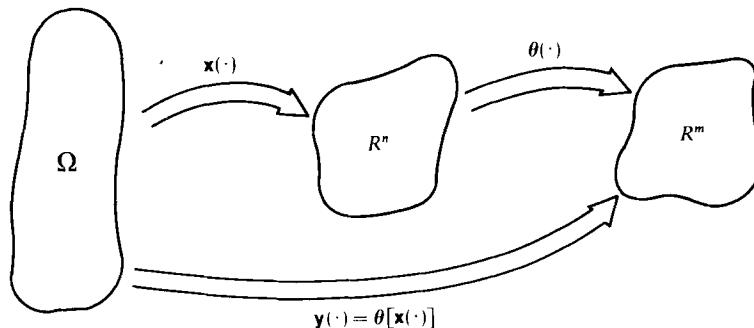
FIG. 3.14   Function of a random variable.

or

$$y_1(\cdot) = \theta_1[x_1(\cdot), x_2(\cdot), \ldots, x_n(\cdot)]$$
$$\vdots \qquad\qquad\qquad\qquad\qquad (3\text{-}52b)$$
$$y_m(\cdot) = \theta_m[x_1(\cdot), x_2(\cdot), \ldots, x_n(\cdot)]$$

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

Recall Eq. (3-28): $x(\cdot)$ generates a new probability space $(R^n, \mathcal{F}_B, P_x)$ from the original probability space $(\Omega, \mathcal{F}, P)$, with P_x defined in (3-26) or (3-27). If $\theta(\cdot)$ 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, $\{x \in R^n : \theta(x) \in B\}$, is an event for which probability has been defined through P_x. If we were to view $(R^n, \mathcal{F}_B, P_x)$ as the underlying probability space, then this just defines $\theta(\cdot)$ 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, $(R^m, \mathcal{F}_B, P_y)$. The sets of interest in R^m will be the elements of the Borel field $\mathcal{F}_B$ associated with R^m, and for all sets $B \subset R^m$ and $B \in \mathcal{F}_B$, we can define probabilities through an appropriate probability function (measure), $P_y(\cdot)$, to be described shortly. Thus,

$$(\Omega, \mathcal{F}, P) \xrightarrow{x(\cdot)} (R^n, \mathcal{F}_B, P_x) \xrightarrow{\theta(\cdot)} (R^m, \mathcal{F}_B, P_y) \qquad (3\text{-}53)$$

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

$$(\Omega, \mathcal{F}, P) \xrightarrow{y(\cdot) = \theta[x(\cdot)]} (R^m, \mathcal{F}_B, P_y) \qquad (3\text{-}54)$$

The random variable y has a *distribution induced by the distribution of* x:

$$F_y(\rho) = P(\{\omega : y(\omega) \le \rho\}) = P(\{\omega : \theta[x(\omega)] \le \rho\})$$
$$= P_x(\{x : \theta(x) \le \rho\}) \qquad (3\text{-}55)$$

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

$$f_y(\rho) = \frac{\partial^m}{\partial \rho_1 \partial \rho_2 \cdots \partial \rho_m} F_y(\rho) \qquad (3\text{-}56)$$

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 $x(\cdot)$ through

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

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 y defined as

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

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

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

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

Thus we obtain the distribution function for y induced by the distribution of x as plotted in Fig. 3.15.   ■
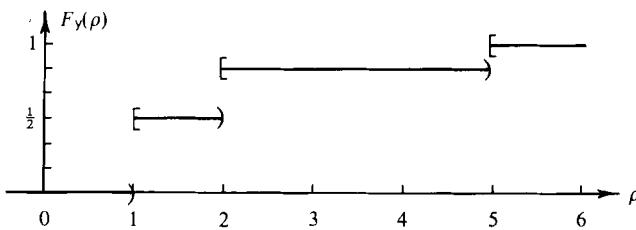
FIG. 3.15   Induced probability distribution function.

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

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

Let the random variable y be defined through

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

Now we want to generate the density function to describe y.

First, the distribution function is

$$F_y(\rho) = P(\{\omega:y(\omega) \le \rho\}) = P(\{\omega:\mathbf{x}^3(\omega) \le \rho\})$$
$$= P_x(\{x:x^3 \le \rho\}) = P_x(\{x:x \le \rho^{1/3}\}) = F_x(\rho^{1/3}) = \int_{-\infty}^{\rho} f_x(\xi)\, d\xi$$

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

$$f_y(\rho) = \frac{dF_y(\rho)}{d\rho} = \frac{d}{d\rho}\left\{ \int_{-\infty}^{\rho^{1/3}} f_x(\xi)\, d\xi \right\}$$
$$= \frac{d}{d\rho}\{\rho^{1/3}\} \cdot f_x(\rho^{1/3}) = \frac{1}{3}\rho^{-2/3} \cdot \frac{1}{\sqrt{2\pi P}} \exp\left\{-\frac{1}{2P}(\rho^{1/3} - m)^2\right\}   ■$$

Thus, given a random variable **x** and its distribution (or density if it exists) and the functional relationship $\mathbf{y} = \theta(\mathbf{x})$, the induced distribution (density) for **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 $\mathbf{y} = \theta(\mathbf{x})$. Suppose θ^{-1} exists and both θ and θ^{-1} are continuously differentiable. Then

$$f_{\mathbf{y}}(\rho) = f_{\mathbf{x}}[\theta^{-1}(\rho)]\|\partial\theta^{-1}(\rho)/\partial\rho\| \qquad (3\text{-}57)$$

where $\|\partial\theta^{-1}(\rho)/\partial\rho\| > 0$ 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

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

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

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

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 $\mathbf{y} = \theta(\mathbf{x})$, where θ maps R^n into R^m, then

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

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 θ. 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 $\hat{\theta}(\cdot)$ from R^m into R^n, called an estimator, that will map the given realizations into a "best" estimate of the value of θ. The composite mapping $\hat{\theta}[\mathbf{z}(\cdot)]$ can then be considered a random

variable, often denoted as $\hat{\theta}$, called a randomized estimate or estimator, or just an estimate. Finally, estimation algorithms generate vectors of numbers, $\hat{\theta}(z) \in R^n$, 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 $\mathbf{x}$:

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

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

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

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

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

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 $f_x(\xi)$ 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

$$E[y] = \int \rho \, dF_y(\rho) = \sum_i \rho_i \Delta F_y(\rho_i) = 1 \cdot (1 - P_1 - P_2) + 2 \cdot P_1 + 5 \cdot P_2$$

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

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,

$$E[c\mathbf{y}] = cE[\mathbf{y}] \qquad (3\text{-}64\text{a})$$

$$E[\mathbf{y}_1 + \mathbf{y}_2] = E[\mathbf{y}_1] + E[\mathbf{y}_2] \qquad (3\text{-}64\text{b})$$

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

$$E[\mathbf{Ay}] = \mathbf{A}E[\mathbf{y}] \qquad (3\text{-}65)$$

Now let us consider some specific functions $\theta(\cdot)$. First let $\theta(\mathbf{x}) = \mathbf{x}$ 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 $m_i \triangleq E[x_i]$:

$$\mathbf{m} \triangleq \begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix} \triangleq \begin{bmatrix} E[x_1] \\ \vdots \\ E[x_n] \end{bmatrix} = \begin{bmatrix} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_1 f_x(\xi) \, d\xi_1 \cdots d\xi_n \\ \vdots \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_n f_x(\xi) \, d\xi_1 \cdots d\xi_n \end{bmatrix} \qquad (3\text{-}66)$$

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

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

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

$$\mathbf{y} = \theta(\mathbf{x}) = \mathbf{x}\mathbf{x}^T = \begin{bmatrix} x_1^2 & x_1 x_2 & \cdots & x_1 x_n \\ \vdots & \vdots & & \vdots \\ x_n x_1 & x_n x_2 & \cdots & x_n^2 \end{bmatrix} \qquad (3\text{-}68)$$

Define a matrix, denoted as $\mathbf{\Psi}$, 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):

$$\Psi_{ij} \triangleq E[x_i x_j] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_i \xi_j f_{\mathbf{x}}(\xi) \, d\xi_1 \cdots d\xi_n \qquad (3\text{-}69)$$

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

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

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

Let us consider yet another function, $\theta(\mathbf{x}) = [(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T]$. 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:

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

Note specifically that the mean values $m_i = E[x_i]$ and $m_j = E[x_j]$ 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

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

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 $\mathbf{\Psi}_{yy}$.

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:

$$P_{ii} \triangleq E[(x_i - m_i)^2] \qquad (3\text{-}73)$$

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

$$P_{ii} \triangleq \sigma_i^2 \qquad (3\text{-}74)$$

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

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

Using (3-74) and (3-75), the covariance matrix **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 **P** is diagonal, i.e., if $r_{ij} = 0$ for all i and j with $i \neq j$, then **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 **m** is not random:

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

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

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

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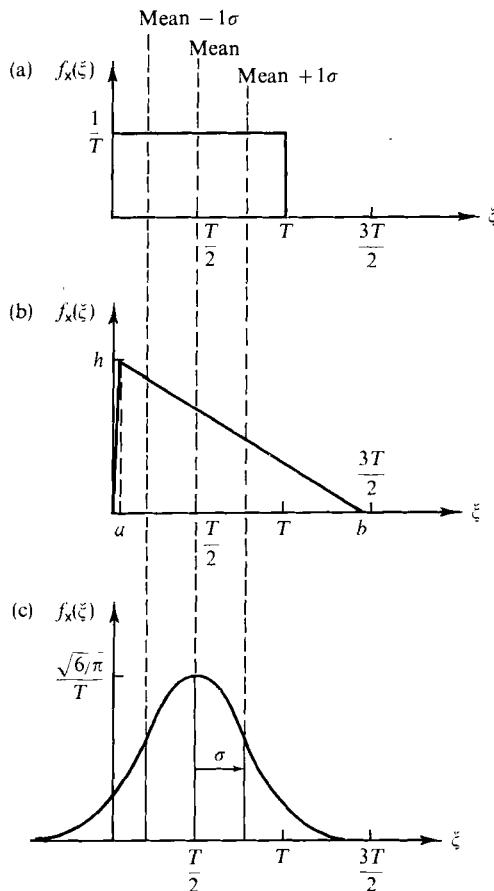
FIG. 3.16   Different random variables with equivalent first two moments. (a) Uniform. (b) Triangular. $a = (T/4)(3 - 5\sqrt{1/3}) \cong 0.03\ T$, $b = (T/4)(3 + 5\sqrt{1/3}) \cong 1.47\ T$, $h = (2/b) \cong 1.36\ T$. (c) Gaussian. $\sigma = T/\sqrt{12} \cong 0.29\ T$.

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

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

and the variance of x is

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

If y is defined as sin x, then $E[y]$ is

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

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 **x** to the second moment relationship between two random variables **x** and **y**. Such a concept is inherently involved in Eqs. (3-70) and (3-72), since the i–j component of **Ψ** or **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 **x** be an n-dimensional random vector and **y** an m-dimensional random vector. Then the *cross-correlation matrix* of **x** and **y** is the n-by-m matrix whose i–j component is

$$E[x_i y_j] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \xi_i \rho_j f_{x,y}(\xi, \rho) d\xi_1 \cdots d\xi_n d\rho_1 \cdots d\rho_m \qquad (3\text{-}79)$$

This matrix is then expressed notationally as

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

Similarly, the second central moment generalizes to the *cross-covariance matrix* **x** and **y**:

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

$$(3\text{-}81)$$

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

$$E[\mathbf{x}\mathbf{y}^T] = E[\mathbf{x}]E[\mathbf{y}^T] = \mathbf{m}_x \mathbf{m}_y^T \qquad (3\text{-}82a)$$

or

$$E[x_i y_j] = E[x_i]E[y_j] \qquad \text{for all} \quad i \text{ and } j \qquad (3\text{-}82b)$$

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

EXAMPLE 3.13 We want to show by a simple example that the preceding definition of un-correlatedness 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 $E[z_1 z_2] = E[z_1]E[z_2]$. 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

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

But, if z_1 and z_2 are uncorrelated, then $E[z_1 z_2] - E[z_1]E[z_2] = 0$, 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

$$\mathbf{x} \text{ and } \mathbf{y} \text{ independent} \rightarrow \mathbf{x} \text{ and } \mathbf{y} \text{ uncorrelated} \qquad (3\text{-}83)$$

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

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

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

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

Separating the integration yields the desired result:

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

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:

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

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 $E[x^2 y^2]$. If x and y were independent, these would be equal. ∎

Another related concept is that of orthogonality. Two random vectors **x** and **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 **x** and **y** being uncorrelated, and this relation is as follows. If either **x** or **y** (or both) is zero-mean, then orthogonality and uncorrelatedness of **x** and **y** imply each other. However, if neither is zero-mean, then **x** and **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 **x** based on measurement data **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 **x** and **y** be random variables mapping Ω into R^n and R^m, respectively, and let **z** be a continuous (Baire) function of **x**,

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

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

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

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

The subscript **x** on $E_{\mathbf{x}}[\mathbf{z}|\mathbf{y} = \mathbf{y}]$ denotes that the expectation operation (integration) is performed over the possible values of **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 **y** values are realizations of the random variable **y**, then the conditional
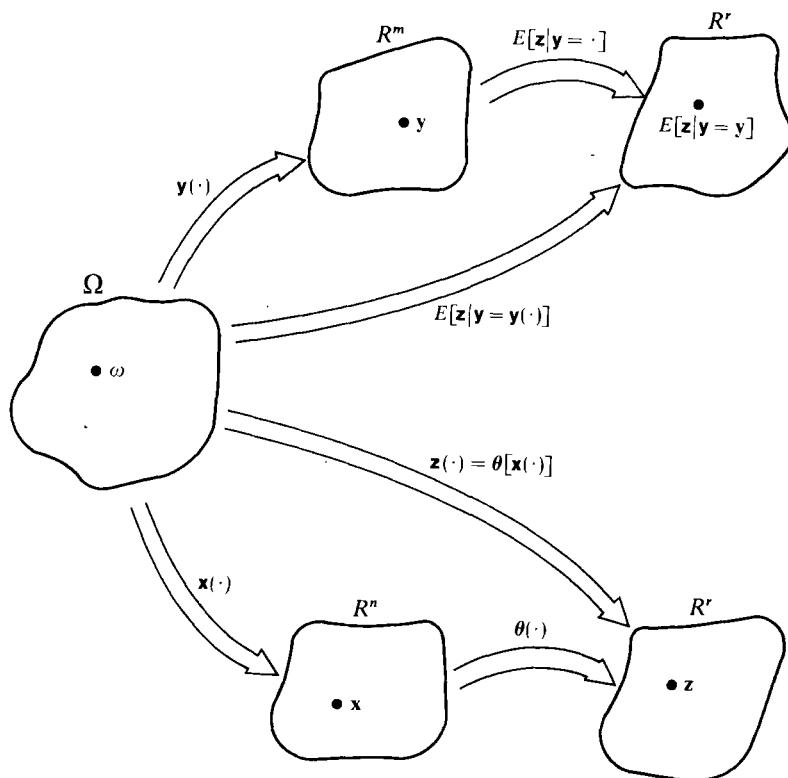
FIG. 3.17   Conditional expectation functional relationships.

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,

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

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

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 **x**.

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

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

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 $\mathscr{F}$, and probability function P; let **x** be a proper random variable (the set $\{\omega:\mathbf{x}(\omega) \leq \xi\}$ is in $\mathscr{F}$). Now let $\mathscr{F}'$ be a σ-algebra that is a subset of $\mathscr{F}$ ($\mathscr{F}'$ is a "coarser" σ-algebra, with fewer or the same number of elements as $\mathscr{F}$). The conditional expectation of **x** relative to $\mathscr{F}'$, $E[\mathbf{x}|\mathscr{F}']$, is any ω function that is a proper random variable relative to $\mathscr{F}'$ (the set $\{\omega:E[\mathbf{x}|\mathscr{F}'] \leq \xi\}$ is in $\mathscr{F}'$; i.e., measurable relative to $\mathscr{F}'$) satisfying

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

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

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

i.e., it is the random variable function of ω, $E\{\mathbf{x}|\mathscr{F}'\}$, 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 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:

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

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, $E\{y|\mathscr{F}'\}$:

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

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

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

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

$$\frac{3}{2} = E\{y|\mathscr{F}'\}\Big|_{\omega \in (A_1 \cup A_2)} \int_{(A_1 \cup A_2)} dP(\omega)$$

$$= E\{y|\mathscr{F}'\}\Big|_{\omega \in (A_1 \cup A_2)} \left[\frac{1}{3} + \frac{1}{6}\right]$$

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

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

Similar reasoning then yields

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

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

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

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

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 $\boldsymbol{\mu} = \mathbf{0}$:

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

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 $E[x_k x_l]$ can be evaluated as

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

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

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

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,

$$\mathbf{z} = \mathbf{x} + \mathbf{y} \qquad (3\text{-}97)$$

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)]:

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

But, since **x** and **y** are independent, this equals the unconditional probability that **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

$$f_{\mathbf{z}}(\zeta) = \int_{-\infty}^{\infty} f_{\mathbf{z},\mathbf{x}}(\zeta,\xi)\,d\xi = \int_{-\infty}^{\infty} f_{\mathbf{z}|\mathbf{x}}(\zeta \,|\, \xi) f_{\mathbf{x}}(\xi)\,d\xi$$

$$= \int_{-\infty}^{\infty} f_{\mathbf{y}}(\zeta - \xi) f_{\mathbf{x}}(\xi)\,d\xi \tag{3-99}$$

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.

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

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,

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

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\boldsymbol{\mu}^{\mathrm{T}}(\xi+\rho)} f_{\mathbf{y}}(\zeta - \xi) f_{\mathbf{x}}(\xi)\,d\zeta \right] d\xi$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\boldsymbol{\mu}^{\mathrm{T}}(\xi+\rho)} f_{\mathbf{y}}(\rho) f_{\mathbf{x}}(\xi)\,d\rho \right] d\xi$$

$$= \left[\int_{-\infty}^{\infty} e^{j\boldsymbol{\mu}^{\mathrm{T}}\xi} f_{\mathbf{x}}(\xi)\,d\xi \right] \left[\int_{-\infty}^{\infty} e^{j\boldsymbol{\mu}^{\mathrm{T}}\rho} f_{\mathbf{y}}(\rho)\,d\rho \right]$$

$$= \phi_{\mathbf{x}}(\boldsymbol{\mu})\phi_{\mathbf{y}}(\boldsymbol{\mu}) \tag{3-100}$$

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

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

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:

$$f_{x}(\xi) = \frac{1}{\sqrt{2\pi P}} \exp\left\{-\frac{1}{2P}(\xi - m)^{2}\right\} \qquad (3\text{-}102)$$

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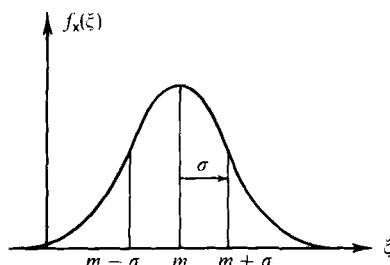
FIG. 3.18   Density function of a scalar Gaussian random variable.

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

$$f_\mathbf{x}(\boldsymbol{\xi}) = (2\pi)^{-1} \left| \begin{bmatrix} \sigma_1{}^2 & r_{12}\sigma_1\sigma_2 \\ r_{12}\sigma_1\sigma_2 & \sigma_2{}^2 \end{bmatrix} \right|^{-1/2}$$

$$\times \exp -\frac{1}{2}\left\{ \begin{bmatrix} \xi_1 - m_1 \\ \xi_2 - m_2 \end{bmatrix}^\mathsf{T} \begin{bmatrix} \sigma_1{}^2 & r_{12}\sigma_1\sigma_2 \\ r_{12}\sigma_1\sigma_2 & \sigma_2{}^2 \end{bmatrix}^{-1} \begin{bmatrix} \xi_1 - m_1 \\ \xi_2 - m_2 \end{bmatrix} \right\}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2(1 - r_{12}^2)^{1/2}} \exp\left\{ -\frac{1}{2(1 - r_{12})^2} \left[\frac{(\xi_1 - m_1)^2}{\sigma_1{}^2} \right.\right.$$

$$\left.\left. + \frac{(\xi_2 - m_2)^2}{\sigma_2{}^2} - \frac{2r_{12}(\xi_1 - m_1)(\xi_2 - m_2)}{\sigma_1\sigma_2} \right] \right\} \qquad (3\text{-}103)$$

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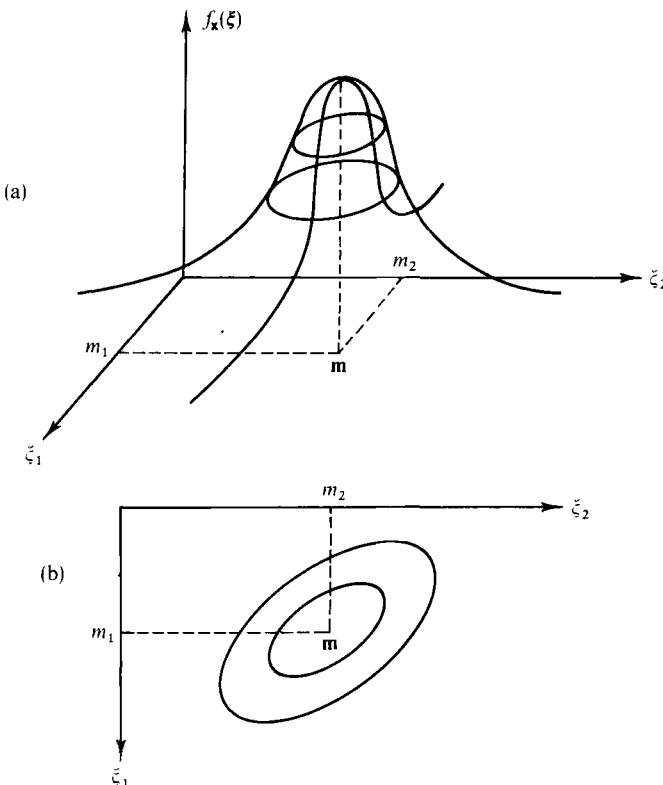
FIG. 3.19   Density function for a two-dimensional Gaussian random vector. (a) Three-dimensional depiction. (b) View from above. The ellipses are the loci of constant probability density value.

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* **x** assumes a value somewhere on the line.

These ideas generalize to higher-dimensional 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 **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

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

Thus, $\phi_x(\mu)$ becomes

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

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 $\mathbf{A}^T = \mathbf{A}^{-1}$ and

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

Thus,

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

Define a new set of variables through a coordinate rotation as

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

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

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

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

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

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

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

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

$$= \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}\beta_i^2/\sigma_i^2 - \tfrac{1}{2}\rho_i^2\sigma_i^2\}\, d\beta_i$$

$$= \exp\{-\tfrac{1}{2}\rho_i^2\sigma_i^2\} \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}\beta_i^2/\sigma_i^2\}\, d\beta_i$$

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:

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

$$= \sqrt{2\pi}\sigma_i \exp\{-\tfrac{1}{2}\rho_i^2\sigma_i^2\}[1]$$

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

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

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 P, $|P| = \sigma_1^2\sigma_2^2 \cdots \sigma_n^2$ and so $|P|^{1/2} = \sigma_1\sigma_2 \cdots \sigma_n$, and this then cancels $\prod_{i=1}^{n} \sigma_i$. Thus,

$$\phi_x(\mu) = \exp\{j\mu^T m\} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \rho_i^2\sigma_i^2 \right\}$$

$$= \exp\{j\mu^T m\} \exp\left\{ -\frac{1}{2}\rho^T \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix} \rho \right\}$$

$$= \exp\{j\mu^T m\} \exp\left\{ -\frac{1}{2}\mu^T A \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix} A^T \mu \right\}$$

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

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

Substituting this into the quadratic form in $\phi_x(\mu)$ yields

$$\phi_x(\mu) = \exp\{j\mu^T m\} \exp\{-\tfrac{1}{2}\mu^T P\mu\}$$

$$= \exp\{j\mu^T m - \tfrac{1}{2}\mu^T P\mu\}$$

as claimed in (3-105).   ■

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

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

P in the density and characteristic functions are the *mean* and *covariance* of **x**, respectively:

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

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

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

DERIVATION OF MEAN AND COVARIANCE   The characteristic function for Gaussian **x** is

$$\phi_{\mathbf{x}}(\mu) = \exp\{j\mu^T\mathbf{m} - \tfrac{1}{2}\mu^T\mathbf{P}\mu\}$$

$$= \exp\left\{ j \sum_{i=1}^{n} \mu_i m_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{q=1}^{n} \mu_i \mu_q P_{iq} \right\}$$

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

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

so we can write

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

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

$$\frac{\partial^2 \phi_{\mathbf{x}}(\mu)}{\partial \mu_k \partial \mu_l} = \left[\frac{\partial}{\partial \mu_l} \left(j m_k - \sum_{q=1}^{n} P_{kq}\mu_q \right) \right] \phi_{\mathbf{x}}(\mu) + \left(j m_k - \sum_{q=1}^{n} P_{kq}\mu_q \right) \frac{\partial \phi_{\mathbf{x}}(\mu)}{\partial \mu_l}$$

$$= (-P_{kl})\phi_{\mathbf{x}}(\mu) + \left(j m_k - \sum_{q=1}^{n} P_{kq}\mu_q \right) \left(j m_l - \sum_{q=1}^{n} P_{lq}\mu_q \right) \phi_{\mathbf{x}}(\mu)$$

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

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

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

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

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

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

$$E[\mathbf{x}_k] = 0 \qquad E[\mathbf{x}_k \mathbf{x}_l \mathbf{x}_m] = 0$$

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

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—**x** and **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 **x** and **y** are random vectors, of dimensions n and m, respectively, that are jointly Gaussian and uncorrelated. Define **z** to be a random vector of dimension $(n + m)$ composed of components **x** and **y**:

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

To say **x** and **y** are jointly Gaussian is equivalent to saying that **z** is Gaussian. The first two moments of **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 **x** and **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 (**x** values) and ρ (**y** values):

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

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

$$f_z(\zeta) = [(2\pi)^{(n+m)/2}|\mathbf{P}_{zz}|^{1/2}]^{-1} \exp\{-\tfrac{1}{2}[\zeta - \mathbf{m}_z]^T\mathbf{P}_{zz}^{-1}[\zeta - \mathbf{m}_z]\}$$

$$= \left[(2\pi)^{(n+m)/2}\left|\begin{bmatrix} \mathbf{P}_{xx} & | & \mathbf{0} \\ \hline \mathbf{0} & | & \mathbf{P}_{yy} \end{bmatrix}\right|^{1/2}\right]^{-1} \exp\left\{-\frac{1}{2}\begin{bmatrix} \xi - \mathbf{m}_x \\ \rho - \mathbf{m}_y \end{bmatrix}^T\begin{bmatrix} \mathbf{P}_{xx}^{-1} & | & \mathbf{0} \\ \hline \mathbf{0} & | & \mathbf{P}_{yy}^{-1} \end{bmatrix}\begin{bmatrix} \xi - \mathbf{m}_x \\ \rho - \mathbf{m}_y \end{bmatrix}\right\}$$

$$= [(2\pi)^{n/2}(2\pi)^{m/2}|\mathbf{P}_{xx}|^{1/2}|\mathbf{P}_{yy}|^{1/2}]^{-1}$$
$$\cdot \exp\{-\tfrac{1}{2}[\xi - \mathbf{m}_x]^T\mathbf{P}_{xx}^{-1}[\xi - \mathbf{m}_x] - \tfrac{1}{2}[\rho - \mathbf{m}_y]^T\mathbf{P}_{yy}^{-1}[\rho - \mathbf{m}_y]\}$$

$$= [(2\pi)^{n/2}|\mathbf{P}_{xx}|^{1/2}]^{-1} \exp\{-\tfrac{1}{2}[\xi - \mathbf{m}_x]^T\mathbf{P}_{xx}^{-1}[\xi - \mathbf{m}_x]\}$$
$$\cdot [(2\pi)^{m/2}|\mathbf{P}_{yy}|^{1/2}]^{-1} \exp\{-\tfrac{1}{2}[\rho - \mathbf{m}_y]^T\mathbf{P}_{yy}^{-1}[\rho - \mathbf{m}_y]\}$$

$$= [f_{\mathbf{x}}(\xi)] \cdot [f_{\mathbf{y}}(\rho)]$$

Thus, **x** and **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.   ∎
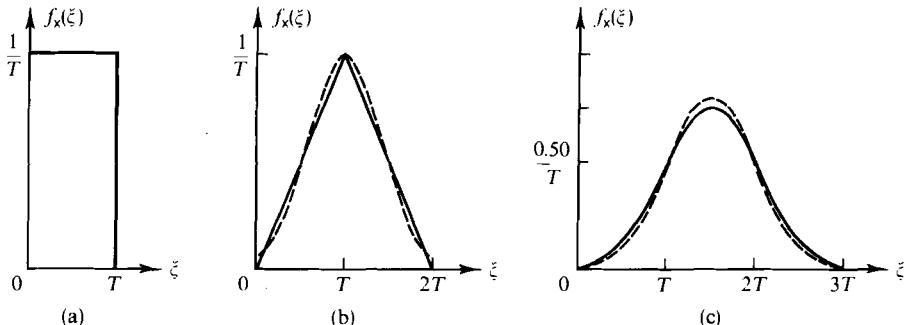
FIG. 3.20   Central limit theorem exemplified. (a) $f_{x_i}(\xi)$. (b) $x = x_1 + x_2$. Solid line indicates $f_x(\xi)$; dashed, $(1/T)\sqrt{3/\pi}\ \exp[-3(x - T)^2/T^2]$. (c) $x = x_1 + x_2 + x_3$. Solid line indicates $f_x(\xi)$; dashed, $(1/T)\sqrt{2/\pi}\ \exp[-2(x - 1.5\ T)^2/T^2]$. From *Probability, 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 **x** and **y** be jointly Gaussian vectors mapping Ω into R^n and R^m, respectively, so that $f_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi},\boldsymbol{\rho})$ can be written as

$$f_{\mathbf{x},\mathbf{y}}(\boldsymbol{\xi},\boldsymbol{\rho}) = \left[(2\pi)^{(n+m)/2} \left| \begin{bmatrix} \mathbf{P}_{xx} & \vline & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & \vline & \mathbf{P}_{yy} \end{bmatrix} \right| \right]^{-1}$$

$$\times \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\xi} - \mathbf{m}_x \\ \boldsymbol{\rho} - \mathbf{m}_y \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{P}_{xx} & \vline & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & \vline & \mathbf{P}_{yy} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\xi} - \mathbf{m}_x \\ \boldsymbol{\rho} - \mathbf{m}_y \end{bmatrix} \right\} \quad (3\text{-}109)$$

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

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

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

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

where

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

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

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

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

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:

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

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

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

As mentioned before, this function is properly defined for $\mathbf{P}_{x|y}$ positive semidefinite, 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 $\mathbf{y}(\cdot)$:

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

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 **x** is a Gaussian random n-vector with mean $\mathbf{m}_x$ and covariance $\mathbf{P}_{xx}$, and **A** is a known $(m \times n)$ matrix (not random), then the random m-vector **y** defined by

$$\mathbf{y} = \mathbf{Ax} \tag{3-117}$$

is Gaussian with mean and covariance given by

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

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

Proof By definition, the characteristic function for **y** is

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

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

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

Thus,

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

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

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

$$\mathbf{z} = \mathbf{Ax} + \mathbf{By} \tag{3-119}$$

is Gaussian, characterized by mean and covariance

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

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

Proof Form the $(n + m)$-dimensional Gaussian random variable **w**

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

which is characterized by mean and covariance

$$\mathbf{m}_w = \begin{bmatrix} \mathbf{m}_x \\ \hline \mathbf{m}_v \end{bmatrix}, \qquad \mathbf{P}_{ww} = \begin{bmatrix} \mathbf{P}_{xx} & | & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & | & \mathbf{P}_{yy} \end{bmatrix}$$

Also form the matrix $\mathbf{C} = [\mathbf{A} \,|\, \mathbf{B}]$. Then $\mathbf{z} = \mathbf{Cw}$, 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

$$\mathbf{z} = \mathbf{Ax} + \mathbf{By} + \mathbf{c} \tag{3-121}$$

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

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

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

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

$$\phi_{\mathbf{z}}(\boldsymbol{\mu}) = E[e^{j\boldsymbol{\mu}^T\mathbf{z}}] = E[e^{j\boldsymbol{\mu}^T(\mathbf{Ax}+\mathbf{By})}e^{j\boldsymbol{\mu}^T\mathbf{c}}] = e^{j\boldsymbol{\mu}^T\mathbf{c}}E[e^{j\boldsymbol{\mu}^T(\mathbf{Ax}+\mathbf{By})}]$$

The proof is then as before, but with the additional $e^{j\boldsymbol{\mu}^T\mathbf{c}}$ 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

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

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

$$\mathbf{x} = \mathbf{Az} = \mathbf{Ix} + \mathbf{0y}$$

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

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

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

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 **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, **z**.

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

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

where **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 **x** as the realizations of a random variable **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

$$\mathbf{z} = \mathbf{Hx} + \mathbf{v} \qquad (3\text{-}126)$$

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 $f_{\mathbf{x}|\mathbf{z}}(\xi\,|\,\mathbf{z})$. 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 $f_{\mathbf{x}|\mathbf{z}}(\xi\,|\,\mathbf{z})$ 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
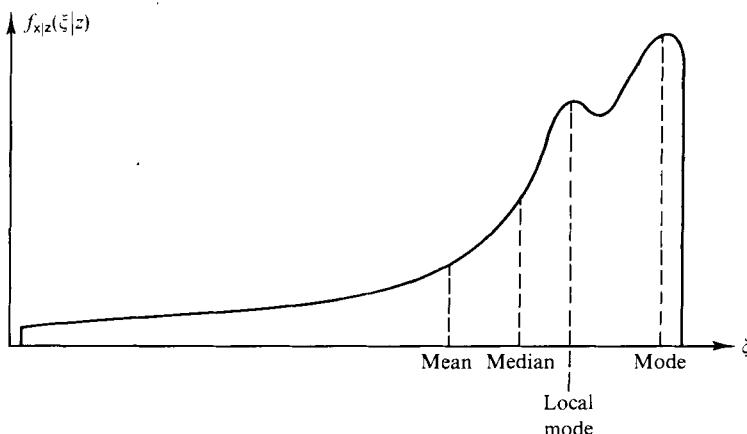
FIG. 3.21   Choice of estimator.

(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

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

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

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

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

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

$$\mathbf{P}_{ww} = \begin{bmatrix} \mathbf{I} & 0 \\ \hline \mathbf{H} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{P}^- & 0 \\ \hline 0 & \mathbf{R} \end{bmatrix}\begin{bmatrix} \mathbf{I} & \mathbf{H}^T \\ \hline 0 & \mathbf{I} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{P}^- & \mathbf{P}^-\mathbf{H}^T \\ \hline \mathbf{H}\mathbf{P}^- & \mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R} \end{bmatrix} \tag{3-131b}$$

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:

$$\text{Random variable:} \quad \mathbf{y} \to \mathbf{z}$$

$$\text{Realization:} \quad \mathbf{y} \to \mathbf{z}$$

$$\text{Dummy variable:} \quad \rho \to \zeta$$

$$\text{Mean of joint density:} \quad \begin{bmatrix} \mathbf{m}_x \\ \hline \mathbf{m}_y \end{bmatrix} \to \begin{bmatrix} \hat{\mathbf{x}}^- \\ \hline \mathbf{H}\hat{\mathbf{x}}^- \end{bmatrix}$$

$$\text{Covariance of joint density:} \quad \begin{bmatrix} \mathbf{P}_{xx} & | & \mathbf{P}_{xy} \\ \hline \mathbf{P}_{yx} & | & \mathbf{P}_{yy} \end{bmatrix} \to \begin{bmatrix} \mathbf{P}^- & | & \mathbf{P}^-\mathbf{H}^T \\ \hline \mathbf{H}\mathbf{P}^- & | & \mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R} \end{bmatrix}$$

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

$$\hat{\mathbf{x}}^+ = E_{\mathbf{x}}[\mathbf{x}|\mathbf{z} = \mathbf{z}] = \hat{\mathbf{x}}^- + [\mathbf{P}^-\mathbf{H}^T][\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]^{-1}[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}^-] \quad (3\text{-}132)$$

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

$$\mathbf{P}^+ = \mathbf{P}^- - [\mathbf{P}^-\mathbf{H}^T][\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]^{-1}[\mathbf{H}\mathbf{P}^-] \quad (3\text{-}133)$$

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

$$\mathbf{K} = \mathbf{P}^-\mathbf{H}^T[\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]^{-1} \quad (3\text{-}134)$$

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

$$\hat{\mathbf{x}}^+ = \hat{\mathbf{x}}^- + \mathbf{K}[\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}^-] \quad (3\text{-}135)$$

$$\mathbf{P}^+ = \mathbf{P}^- - \mathbf{K}\mathbf{H}\mathbf{P}^- \quad (3\text{-}136)$$

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

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

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

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

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

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}^+ \quad (3\text{-}139)$$

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

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

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

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

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

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 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 $z_1(\omega) = z_1$ 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 $\sigma_{z_1}^2$ ("$\hat{x}^-$" $= z_1$, "P^-" $= \sigma_{z_1}^2$). Consequently, we consider $z_2(\omega) = z_2$ 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 "R" $= \sigma_{z_2}^2$.

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

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

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^+:

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

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

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

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

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

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

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

$$= \left\{ \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma_{z_1}^2 & 0 \\ 0 & 1/\sigma_{z_2}^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}^{-1} = \frac{1}{[(1/\sigma_{z_1}^2) + (1/\sigma_{z_2}^2)]}$$

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):

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

$$= P^+\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{z} = [\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{z}$$

$$= \frac{\sigma_{z_1}^2 \sigma_{z_2}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma_{z_1}^2 & 0 \\ 0 & 1/\sigma_{z_2}^2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$= \frac{\sigma_{z_2}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_1 + \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{z_2}^2} z_2$$

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 z be partitioned corresponding to the diagonal blocks of **R**:

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

Then batch processing of 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 **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 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 **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 z to generate an estimate $\hat{\mathbf{x}}$ of the unknown 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 **W** be a general $(m \times m)$ weighting matrix, then we

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

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

Note that if $W = I$, this is standard least squares, with

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

If 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 $\hat{x}_{WLS}$, where the subscript denotes weighted least squares, if

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

and

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

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

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

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

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

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

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

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

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

The two results are identical if we choose W to be R^{-1} (positive definite). However, least squares theory gives *no* insights into such a choice of weighting matrix, since no statistical characterization (as v being Gaussian, zero-mean, of covariance R) was assumed to be known. Analogous to the previous discussion, if 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 $x(\cdot)$ to consider the number of heads appearing in two tosses. Obtain the probability distribution function for this $x(\cdot)$.

3.3 If the joint probability density of x_1 and x_2 is

$$
f_{x_1, x_2}(\xi_1, \xi_2) = \begin{cases} \dfrac{1}{2\pi} \exp\left[-\tfrac{1}{2}(\xi_1 - 3)^2 - \xi_2\right] & \xi_2 > 0 \\[2mm] 0 & \xi_2 \leq 0 \end{cases}
$$

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

$$
y = x_1 + x_2 \ ?
$$

Determine the mean of y.

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

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

Is this the same as

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

Show why.

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

$$
z(\cdot) = x(\cdot)y(\cdot)
$$

(a) What are the mean, mean squared value, and variance of $x(\cdot)$ or $y(\cdot)$? For $z(\cdot)$?

(b) What is the probability that $z(\cdot)$ 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 $\mathbf{y} = \theta(\mathbf{x})$. Suppose θ^{-1} exists and that both θ and θ^{-1} are continuously differentiable. Then

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

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

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

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

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

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

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

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

   (a)   Find the probability density for $\mathbf{y}(\cdot)$ by using fundamental set definitions to establish $F_\mathbf{y}(\rho)$ and then find its derivative.
   (b)   Find $f_\mathbf{y}(\rho)$ by a method analogous to the proof of the preceding problem.
   (c)   Find $f_\mathbf{y}(\rho)$ by direct application of the result of the last problem.
   (d)   Find the probability density for $\mathbf{z}(\cdot)$ 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 $[x, y, z]^\mathrm{T}$ is a Gaussian random vector?

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

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

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.,

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

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

$$x_1'' = x_1'/\sigma_{11}', \qquad x_2'' = x_2'/\sigma_{22}', \qquad x_3'' = x_3'/\sigma_{33}'$$

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 $\mathbf{P}$,

$$E[x_k x_l x_m x_n] = P_{kl}P_{mn} + P_{km}P_{ln} + P_{kn}P_{lm}$$

3.13   At the end of Section 3.9, it was claimed that the error $\{\mathbf{x} - E_x[\mathbf{x}|\mathbf{y} = \mathbf{y}(\cdot)]\}$ 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

$$z = x + n$$

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 $f_{x|z}(\xi|\zeta)$ for $\zeta = \frac{1}{2}$, i.e., $f_{x|z}(\xi|\frac{1}{2})$. Plot this density as a function of ξ.

One reasonable estimate of x would be the value of ξ that maximizes the density $f_{x|z}(\xi|\frac{1}{2})$. 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, $E_x[x|z = \frac{1}{2}]$, 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

$$e_{\text{ML}} = \hat{x}_{\text{ML}} - x$$
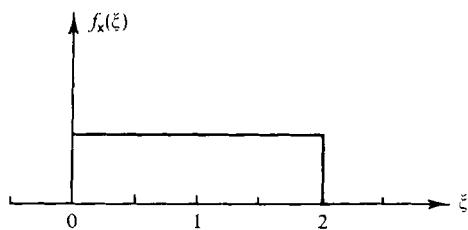
FIG. 3.P1   A priori information for Problem 3.14.
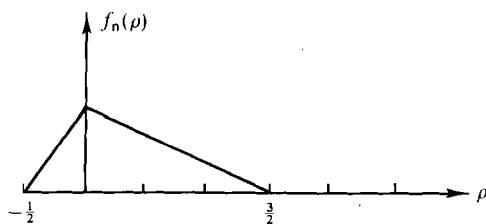
FIG. 3.P2   Measurement noise description for Problem 3.14.

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.
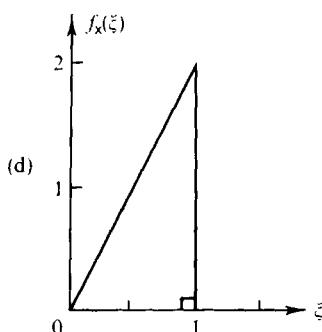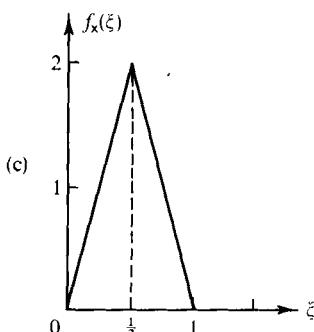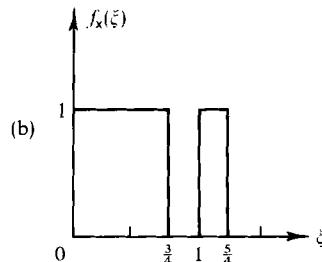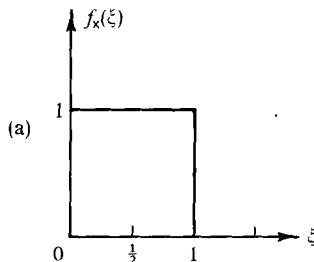
FIG. 3.P3   Density functions for Problem 3.15.

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 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 $E_x[\mathbf{x}|\mathbf{z} = \mathbf{z}]$? 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 **x**, conditioned on the fact that a random variable **z**, jointly Gaussian with **x**, assumed some value, 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 $E[\mathbf{x}|\mathbf{z} = z]$ as an *estimate* of the value of x. Explain this new interpretation of $\mathbf{P}_{x|z}$ explicitly. Show that it is a valid interpretation by showing that if **y** is a continuous (Baire) function of **z**, $\mathbf{y} = \boldsymbol{\theta}(\mathbf{z})$, then

$$E\{\mathbf{xy}^\mathbf{T}\,|\,\mathbf{z}(\omega_j) = z\} = E\{\mathbf{x}\,|\,\mathbf{z}(\omega_j) = z\}\,[\boldsymbol{\theta}^\mathbf{T}(z)]$$

and use this result to prove (3-140).

Is $\mathbf{P}_{x|z}$ a function of the value z that **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 $E[xz_1] \neq 0$, then the value of z_1, i.e., z_1, is always a part of the best estimate of $x(\omega_k) = x$.

(b)   If $E[xz_1] = 0$, then the value of $z_1(z_1)$ 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 z assumed by the measurements **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

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

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 $x(\cdot)$, 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 $x(\cdot)$ 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 $x_1(\cdot), x_2(\cdot), \ldots, x_N(\cdot)$ be N random variables, each with distribution identical to that of $x(\cdot)$. Thus, we can generate the estimator

$$\hat{M}_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and conceive of conducting an experiment of generating the N data points repeatedly, the jth such experiment yielding a single realization of $\hat{M}_x$, $\hat{M}_x(\omega_j)$. We want to characterize the distribution of these estimate values.

(a) Show that $E\{\hat{M}_x\} = \mu_x$: that $\hat{M}_x$ is an unbiased estimator of the mean of $x(\cdot)$.

(b) Show that the variance of $\hat{M}_x$, denoted as $\sigma_{\hat{m}}^2$, is

$$\sigma_{\hat{m}}^2 = \frac{1}{N} E\{x^2\} + \frac{2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E\{x_i x_j\} - \mu_x^2$$

so that, if the observations are independent of each other (i.e., $x_1, x_2, \ldots, x_N$ 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, $\hat{M}_x$ becomes more and more Gaussian, regardless of the distribution of x. (Why?) If $\hat{M}_x$ 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 $x(\cdot)$ is defined as

$$\sigma_x^2 \triangleq E\{[x - E\{x\}]^2\} = E\{x^2\} - [E\{x\}]^2$$

a reasonable estimator of the variance of $x(\cdot)$ 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

$$E\{\hat{V}_x'\} = \sigma_x^2 - \sigma_{\hat{m}}^2 \neq \sigma_x^2$$

with $\sigma_{\hat{m}}^2$ defined in the previous problem. For independent observations, show that this becomes

$$E\{\hat{V}_x'\} = [(N - 1)/N]\sigma_x^2$$

(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 $x(\cdot)$, 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

$$\sigma_{\hat{v}}^2 = (1/N)\big[E\{(x - \mu_x)^4\} - \{(N-3)/(N-1)\}\sigma_x^4\big]$$

If $x(\cdot)$ were assumed Gaussian, show that the fourth central moment is equal to $3\sigma_x^4$, and thus

$$\sigma_{\hat{v}}^2 = [2/(N-1)]\sigma_x^4 \qquad [\text{if } x(\cdot) \text{ is Gaussian}]$$

If $x(\cdot)$ were instead assumed to be uniform, show that the fourth central moment is equal to $\frac{9}{5}\sigma_x^4$, and so

$$\sigma_{\hat{v}}^2 = [(4N+6)/\{5N(N-1)\}]\sigma_x^4 \qquad [\text{if } x(\cdot) \text{ is uniform}]$$

Thus, under very different assumptions about $x(\cdot)$, the quality of the estimate provided by $\hat{V}_x$ is very similar:

$$\sigma_{\hat{v}} \ (x \text{ Gaussian}) \cong 1.4\sigma_{\hat{v}} \quad (x \text{ uniform})$$

3.26   Analogous to the variance estimator of the previous problem, a good estimate of the covariance between $x(\cdot)$ and $y(\cdot)$ 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:

$$E\{\hat{C}_{xy}\} = \sigma_{xy}^2 = \text{true covariance of } x(\cdot) \text{ and } y(\cdot)$$

(b)   For independent measurements, the variance of $\hat{C}_{xy}$ is

$$\sigma_{\hat{c}}^2 = (1/N)\big[E\{[x - \mu_x]^2[y - \mu_y]^2\} + [1/(N-1)]\{\sigma_x^2\sigma_y^2 - (N-2)\sigma_{xy}^4\}\big]$$

Show that this measure of the quality of the $\hat{C}_{xy}$ estimate becomes, if $x(\cdot)$ and $y(\cdot)$ 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 $x(\cdot)$ and $y(\cdot)$.

(c)   From the definition of correlation coefficient, (3-75), a good estimator of the (linear) correlation coefficient can be produced as

$$\hat{r}_{xy} = \frac{\hat{C}_{xy}}{(\hat{V}_x \hat{V}_y)^{1/2}} = \frac{N \sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{([N \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2][N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2])^{1/2}}$$
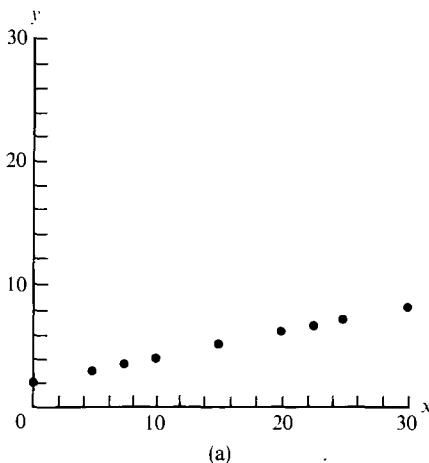
A "scatter diagram" is a two-dimensional plot of the N realizations $[x(\omega_1), y(\omega_1)]$, $[x(\omega_2)y(\omega_2)]$, ...,
$[x(\omega_N), y(\omega_N)]$, as shown in Fig. 3.P4. Plot (a) shows perfect correlation between the two variables,
(b) portrays smaller positive correlation, (c) depicts no correlation, and (d) shows negative correla-
tion. 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
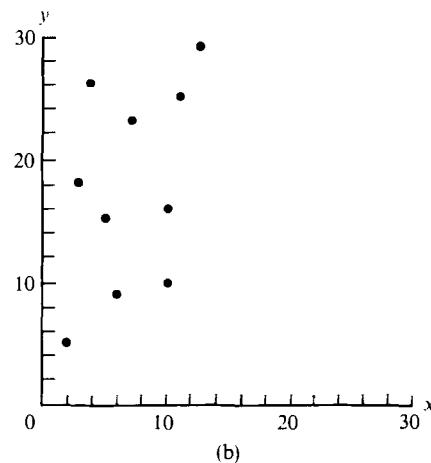
$$y - \hat{M}_y = [\hat{C}_{xy}/\hat{V}_x](x - \hat{M}_x)$$

On the other hand, the least squares regression line of x on y is
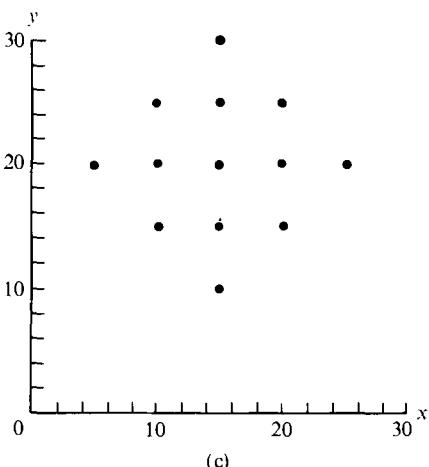
$$x - \hat{M}_x = [\hat{C}_{xy}/\hat{V}_y](y - \hat{M}_y)$$
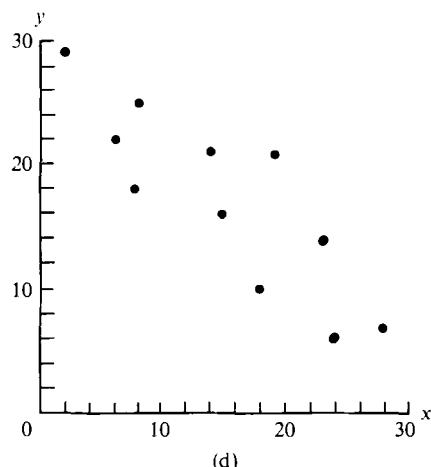
FIG. 3.P4   Scatter diagrams for Problem 3.26.

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

$$y(\cdot) = x(\cdot) + w(\cdot)$$

where $w(\cdot)$ is zero mean, of variance σ_w^2, and independent of $x(\cdot)$. Assume independent observations.

(a) Show that $\hat{M}_x = (1/N)\sum_{i=1}^{N} y_i$ is still an unbiased mean estimator, but with increased variance: $\sigma_{\tilde{m}}^2 = (1/N)(\sigma_x^2 + \sigma_w^2)$.

(b) Show that $\hat{V}_x = [1/(N-1)]\sum_{i=1}^{N} y_i^2 - [N/(N-1)]\hat{M}_x^2$, however, is a biased estimator: $E\{\hat{V}_x\} = \sigma_x^2 + \sigma_w^2$; thus, the best estimator would be $[\hat{V}_x - \sigma_w^2]$. How would σ_w^2 be estimated? If x and w are assumed Gaussian, show that

$$\sigma_{\hat{v}} \text{ (with meas. noise)} = [1 + (\sigma_w^2/\sigma_x^2)]\sigma_{\hat{v}} \text{ (without meas. noise)}.$$

Stochastic processes and linear dynamic system models

4.1 INTRODUCTION

This chapter adds dynamics to the system model developed in Chapter 3, thereby allowing consideration of a much larger class of problems of interest. First, Sections 4.2 and 4.3 characterize stochastic processes in general. Section 4.4 presents the motivation and conceptual framework for developing stochastic linear dynamic system models. State equations in the form of linear stochastic differential or difference equations are developed in Sections 4.5–4.9, and 4.10 adds the description of measured outputs to complete the overall system model. Finally, Sections 4.11–4.13 develop practical system models designed to duplicate (to the extent possible) the characteristics of processes observed empirically.

4.2 STOCHASTIC PROCESSES

Let Ω be a fundamental sample space and T be a subset of the real line denoting a time set of interest. Then a *stochastic process* [1–14] can be defined as a real-valued function $\mathbf{x}(\cdot,\cdot)$ defined on the product space $T \times \Omega$ (i.e., a function of two arguments, the first of which is an element of T and the second an element of Ω), such that for any fixed $t \in T$, $\mathbf{x}(t,\cdot)$ is a random variable. A scalar random process assumes values $\mathbf{x}(t,\omega) \in R^1$, whereas a vector random process assumes values $\mathbf{x}(t,\omega) \in R^n$. In other words, $\mathbf{x}(\cdot,\cdot)$ is a stochastic process if all sets of the form

$$A = \{\omega \in \Omega : \mathbf{x}(t,\omega) \leq \xi\} \tag{4-1}$$

for any $t \in T$ and $\xi \in R^n$ (R^1 for a scalar random process) are in the underlying σ-algebra $\mathscr{F}$. If we fix the second argument instead of the first, we can say that to each point $\omega_i \in \Omega$ there can be associated a time function $\mathbf{x}(\cdot, \omega_i) = \mathbf{x}(\cdot)$, each of which is a *sample* from the stochastic process.

Although the definition of a stochastic process can be generalized to T being a subset of R^n (as for a process as a function of spatial coordinates), we will be interested in $T \subset R^1$ with elements of T being time instants. Two particular forms of T will be important. If T is a sequence $\{t_1, t_2, t_3, \ldots\}$, not necessarily equally spaced, then $\mathbf{x}(t_1, \cdot), \mathbf{x}(t_2, \cdot), \mathbf{x}(t_3, \cdot), \ldots$ becomes a sequence of random variables. This $\mathbf{x}(\cdot, \cdot)$ is then called a discrete-parameter stochastic process, or a *discrete-time stochastic process*. A sample from such a process is depicted in Fig. 4.1; different ω values then generate different samples from the process. If T is instead an interval of R^1, then $\mathbf{x}(\cdot, \cdot)$ becomes a continuous-parameter family of random variables, or a *continuous-time stochastic process*. For each ω, the sample is a function defined on the interval T, as portrayed in Fig. 4.2.

If T is of the discrete form of a finite sequence of N points along the real line, the set of random variables $\mathbf{x}(t_1, \cdot), \mathbf{x}(t_2, \cdot), \ldots, \mathbf{x}(t_N, \cdot)$ can be characterized by
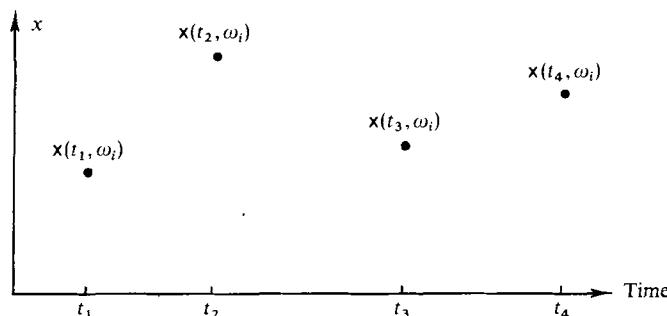
FIG. 4.1   Sample from a discrete-time stochastic process.
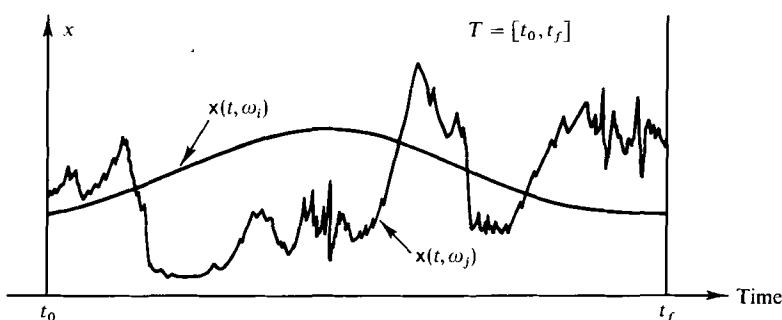
FIG. 4.2   Samples from a continuous-time stochastic process.

the joint probability distribution function

$$F_{\mathbf{x}(t_1),\ldots,\mathbf{x}(t_N)}(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_N) \triangleq P(\{\omega:\mathbf{x}(t_1,\omega)\le\boldsymbol{\xi}_1,\ldots,\mathbf{x}(t_N,\omega)\le\boldsymbol{\xi}_N\}) \quad (4\text{-}2)$$

or the joint density function (if it exists):

$$f_{\mathbf{x}(t_1),\ldots,\mathbf{x}(t_N)}(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_N) = \frac{\partial^{Nn} F_{\mathbf{x}(t_1),\ldots,\mathbf{x}(t_N)}(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_N)}{\partial\xi_{11}\cdots\partial\xi_{1n}\cdots\partial\xi_{N1}\cdots\partial\xi_{Nn}} \quad (4\text{-}3)$$

That is to say, knowledge of such a joint distribution function or the associated joint density function completely describes the set of random variables.

If we want to characterize a continuous-time stochastic process completely, we would require knowledge of the joint probability distribution function $F_{\mathbf{x}(t_1),\ldots,\mathbf{x}(t_N)}(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_N)$ for *all* possible sequences $\{t_1,t_2,\ldots\}$. Again, if it exists, the associated set of joint density functions for all possible time sequences would provide the same complete description. Consider Fig. 4.3. The distribution function $F_{\mathbf{x}(t_1)}(\boldsymbol{\xi}_1)$ establishes the probability of the set of $\omega\in\Omega$ that gives rise to random process samples that assume values less than or equal to $\boldsymbol{\xi}_1$ at time t_1. Similarly, $f_{\mathbf{x}(t_1)}(\boldsymbol{\xi}_1)$ reveals the probability of the set of samples, out of the entire ensemble of process samples, that assume values between $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_1 + d\boldsymbol{\xi}_1$ at time t_1. The joint distribution $F_{\mathbf{x}(t_1),\mathbf{x}(t_2)}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ indicates the probability of the set of samples that not only take on values less than or equal to
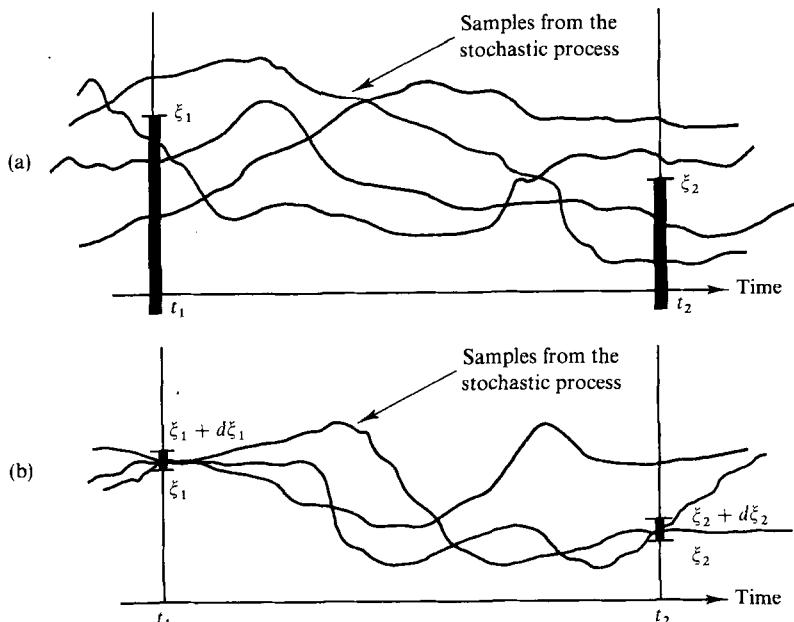
FIG. 4.3 Characterization of stochastic processes. (a) Joint distribution functions. (b) Joint density functions.

ξ_1 at t_1, but also take on values less than or equal to ξ_2 at t_2, and similarly for $f_{\mathbf{x}(t_1),\mathbf{x}(t_2)}(\xi_1,\xi_2)$. These functions more fully characterize the manner in which process samples can change values over time, but even a specification of such functions for all possible t_1 and t_2 would not provide a complete description. That would require an evaluation of all first and second order functions as described, plus all higher order functions for all choices of $t_1, t_2, \ldots$, etc.: conceptually satisfying but infeasible practically.

As can be seen from the definition of a stochastic process and the preceding discussion, the concepts and tools of probability theory from the previous chapter will readily apply to an investigation of stochastic processes. This is particularly true with respect to gaining at least a partial description of a sto-chastic process through a finite number of moments. Rather than trying to generate explicit relations for joint distributions (or densities if they exist), it is often convenient, especially computationally, to describe these functions to some degree by the associated first two moments. In the case of Gaussian processes, such information will completely characterize the joint distribution or density functions, and thus completely characterize the process itself. Note that in the following descriptions the ω argument will be deleted as in the previous chapter, but the boldface sans serif typeface will be maintained to demark stochastic processes: $\mathbf{x}(\cdot,\cdot)$ will be written as $\mathbf{x}(\cdot)$, $\mathbf{x}(t,\cdot)$ becomes the random variable $\mathbf{x}(t)$, and $\mathbf{x}(t,\omega_i) = \mathbf{x}(t)$ will be written as $\mathbf{x}(t)$.

The *mean value function* or mean $\mathbf{m}_x(\cdot)$ of the process $\mathbf{x}(\cdot)$ is defined for all $t \in T$ by

$$\mathbf{m}_x(t) \triangleq E\{\mathbf{x}(t)\} \tag{4-4}$$

i.e., the average value $\mathbf{x}(\cdot)$ assumes at time t, where the average is taken over the entire ensemble of samples from the process. An indication of the spread of values about the mean at time t, $\mathbf{m}_x(t)$, is provided by the second central moment, or *covariance matrix*, $\mathbf{P}_{xx}(\cdot)$, defined by

$$\mathbf{P}_{xx}(t) \triangleq E\{[\mathbf{x}(t) - \mathbf{m}_x(t)][\mathbf{x}(t) - \mathbf{m}_x(t)]^T\} \tag{4-5}$$

A useful generalization of this, containing additional information about how fast $\mathbf{x}(t)$ sample values can change in time, is the *covariance kernel* $\mathbf{P}_{xx}(\cdot,\cdot)$, defined for all $t_1, t_2 \in T$ as

$$\mathbf{P}_{xx}(t_1,t_2) \triangleq E\{[\mathbf{x}(t_1) - \mathbf{m}_x(t_1)][\mathbf{x}(t_2) - \mathbf{m}_x(t_2)]^T\} \tag{4-6}$$

The nature of the information embodied in (4-6) that is not available in (4-5) will be made more explicit in the example to follow. From (4-5) and (4-6), the covariance matrix can be defined as

$$\mathbf{P}_{xx}(t) = \mathbf{P}_{xx}(t,t) \tag{4-7}$$

i.e., $\mathbf{P}_{xx}(\cdot, \cdot)$ with both arguments the same time. The second noncentral moment concept generalizes to the *correlation kernel* $\mathbf{\Psi}_{xx}(\cdot, \cdot)$ defined for all $t_1, t_2 \in T$ as

$$\mathbf{\Psi}_{xx}(t_1, t_2) \triangleq E\{\mathbf{x}(t_1)\mathbf{x}(t_2)^{\mathrm{T}}\} \tag{4-8}$$

The *correlation matrix* would then be $\mathbf{\Psi}_{xx}(t, t)$, composed of individual correlations of components of $\mathbf{x}(t)$: its i–j component would be $E\{x_i(t)x_j(t)\}$, and the diagonal is made up of mean squared values of individual random variables $x_i(t)$. From (4-6) and (4-8), it can be seen that

$$\mathbf{\Psi}_{xx}(t_1, t_2) = \mathbf{P}_{xx}(t_1, t_2) + \mathbf{m}_x(t_1)\mathbf{m}_x(t_2)^{\mathrm{T}} \tag{4-9}$$

and thus if $\mathbf{x}(\cdot)$ is a zero-mean process, $\mathbf{\Psi}_{xx}(t_1, t_2) = \mathbf{P}_{xx}(t_1, t_2)$.

   EXAMPLE 4.1   Consider two scalar zero-mean processes $x(\cdot)$ and $y(\cdot)$ with

$$\Psi_{xx}(t_1, t_2) = P_{xx}(t_1, t_2) = \sigma^2 e^{-|t_1 - t_2|/T}, \qquad \Psi_{yy}(t_1, t_2) = P_{yy}(t_1, t_2) = \sigma^2 e^{-|t_1 - t_2|/10T}$$

where these two correlations are plotted as a function of the time difference $(t_1 - t_2)$ in Fig. 4.4. For a given value of $(t_1 - t_2) \neq 0$, there is a higher correlation between the values of $y(t_1)$ and $y(t_2)$ than between $x(t_1)$ and $x(t_2)$. Physically one would then expect a typical sample $x(\cdot, \omega_i)$ to exhibit more rapid variations in magnitude than $y(\cdot, \omega_i)$, as also depicted in Fig. 4.4. Note that such information is not contained in $P_{xx}(t)$ and $P_{yy}(t)$, or $\Psi_{xx}(t)$ and $\Psi_{yy}(t)$, all of which are the same value for this example, σ^2, as seen by evaluating the preceding expressions for $t_1 = t_2$. ∎
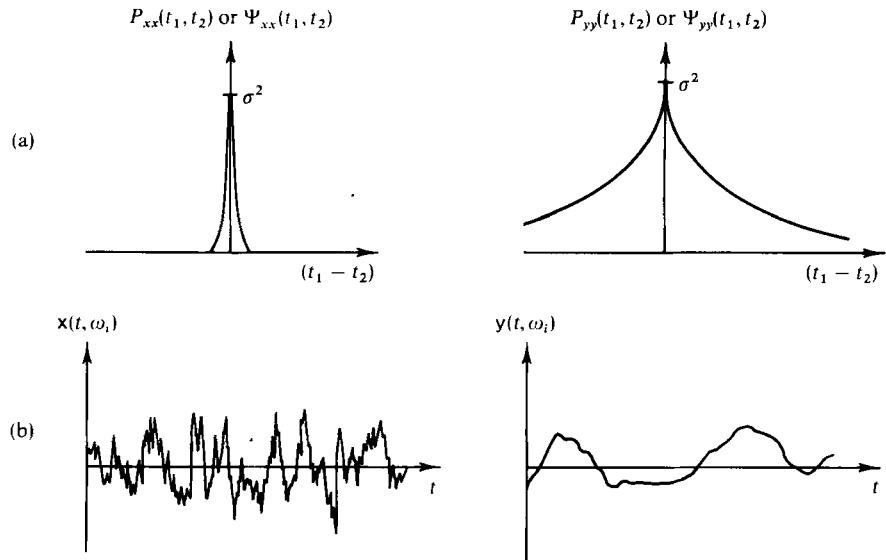
FIG. 4.4   Second moment information about stochastic processes. (a) Correlation or variance kernels. (b) Typical samples from the stochastic processes.

   For characterizing the interrelationship between two stochastic processes $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$, the preceding second moment concepts generalize to the *cross-covariance kernel* of $\mathbf{x}(\cdot)$ and $\mathbf{y}(\cdot)$, $\mathbf{P}_{xy}(\cdot, \cdot)$ defined for all $t_1, t_2 \in T$ as

$$\mathbf{P}_{xy}(t_1, t_2) \triangleq E\{[\mathbf{x}(t_1) - \mathbf{m}_x(t_1)][\mathbf{y}(t_2) - \mathbf{m}_y(t_2)]^{\mathrm{T}}\} \tag{4-10}$$

the *cross-covariance matrix*:

$$\mathbf{P}_{xy}(t) = \mathbf{P}_{xy}(t, t) \tag{4-11}$$

and the *cross-correlation kernel* and associated matrix:

$$\mathbf{\Psi}_{xy}(t_1, t_2) \triangleq E\{\mathbf{x}(t_1)\mathbf{y}(t_2)^\mathrm{T}\} \tag{4-12}$$

$$= \mathbf{P}_{xy}(t_1, t_2) + \mathbf{m}_x(t_1)\mathbf{m}_y{}^\mathrm{T}(t_2) \tag{4-13}$$

Other concepts also readily translate from probability theory, but care must be taken to avoid such ambiguities as the meaning of "independent processes" and "uncorrelated processes." A process $\mathbf{x}(\cdot, \cdot)$ is *independent (in time)* or *white* if, for any choice of $t_1, \ldots, t_N \in T$, $\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)$ are a set of independent random vectors; i.e.,

$$P(\{\omega : \mathbf{x}(t_1, \omega) \le \boldsymbol{\xi}_1, \ldots, \mathbf{x}(t_N, \omega) \le \boldsymbol{\xi}_N\}) = \prod_{i=1}^{N} P(\{\omega : \mathbf{x}(t_i, \omega) \le \boldsymbol{\xi}_i\}) \tag{4-14}$$

or equivalently,

$$F_{\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) = \prod_{i=1}^{N} F_{\mathbf{x}(t_i)}(\boldsymbol{\xi}_i) \tag{4-15}$$

or, if the densities exist,

$$f_{\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) = \prod_{i=1}^{N} f_{\mathbf{x}(t_i)}(\boldsymbol{\xi}_i) \tag{4-16}$$

On the other hand, two processes $\mathbf{x}(\cdot, \cdot)$ and $\mathbf{y}(\cdot, \cdot)$ are said to be *independent (of each other)* if, for any $t_1, \ldots, t_N \in T$,

$$P(\{\omega : \mathbf{x}(t_1, \omega) \le \boldsymbol{\xi}_1, \ldots, \mathbf{x}(t_N, \omega) \le \boldsymbol{\xi}_N, \mathbf{y}(t_1, \omega) \le \boldsymbol{\rho}_1, \ldots, \mathbf{y}(t_N, \omega) \le \boldsymbol{\rho}_N\})$$
$$= P(\{\omega : \mathbf{x}(t_1, \omega) \le \boldsymbol{\xi}_1, \ldots, \mathbf{x}(t_N, \omega) \le \boldsymbol{\xi}_N\})$$
$$\cdot P(\{\omega : \mathbf{y}(t_1, \omega) \le \boldsymbol{\rho}_1, \ldots, \mathbf{y}(t_N, \omega) \le \boldsymbol{\rho}_N\}) \tag{4-17}$$

Thus, "two independent processes" could mean two processes, each of which were independent in time, or two processes independent of each other, or some combination of these. We will use the term "white" to clarify this issue.

In a similar manner, a process $\mathbf{x}(\cdot, \cdot)$ is *uncorrelated (in time)* if, for all $t_1, t_2 \in T$ except for $t_1 = t_2$,

$$\mathbf{\Psi}_{xx}(t_1, t_2) \triangleq E[\mathbf{x}(t_1)\mathbf{x}^\mathrm{T}(t_2)] = E[\mathbf{x}(t_1)]E[\mathbf{x}^\mathrm{T}(t_2)] \tag{4-18a}$$

or

$$\mathbf{P}_{xx}(t_1, t_2) = \mathbf{0} \tag{4-18b}$$

By comparison, two processes $\mathbf{x}(\cdot, \cdot)$ and $\mathbf{y}(\cdot, \cdot)$ are *uncorrelated* with each other if, for all $t_1, t_2 \in T$ (including $t_1 = t_2$),

$$\mathbf{\Psi}_{xy}(t_1, t_2) \triangleq E[\mathbf{x}(t_1)\mathbf{y}^\mathrm{T}(t_2)] = E[\mathbf{x}(t_1)]E[\mathbf{y}^\mathrm{T}(t_2)] \tag{4-19a}$$

or

$$\mathbf{P}_{xy}(t_1, t_2) = \mathbf{0} \qquad (4\text{-}19\text{b})$$

As shown previously, independence implies uncorrelatedness (which restricts attention to only the second moments), but the opposite implication is not true, except in such special cases as Gaussian processes, to be discussed. Note that "white" is often accepted to mean uncorrelated in time rather than independent in time; the distinction between these definitions disappears for the important case of white Gaussian processes.

In the previous chapter, much attention was devoted to Gaussian random variables, motivated by both practical justification (central limit theorem) and mathematical considerations (such as the first two moments completely describing the distribution and Gaussianness being preserved through linear operations). Similarly, Gaussian processes will be of primary interest here. A process $\mathbf{x}(\cdot, \cdot)$ is a *Gaussian process* if all finite joint distribution functions for $\mathbf{x}(t_1, \cdot)$, $\mathbf{x}(t_2, \cdot)$, ..., $\mathbf{x}(t_N, \cdot)$ are Gaussian for any choice of $t_1, t_2, \ldots, t_N$. For instance, if $\mathbf{x}(\cdot, \cdot)$ is Gaussian and the appropriate densities exist, then for any choice of $t_1, t_2 \in T$,

$$F_{\mathbf{x}(t_1), \mathbf{x}(t_2)}(\boldsymbol{\xi}) = \left[(2\pi)^n |\mathbf{P}|^{1/2}\right]^{-1} \exp\left\{-\tfrac{1}{2}(\boldsymbol{\xi} - \mathbf{m})^{\mathrm{T}} \mathbf{P}^{-1}(\boldsymbol{\xi} - \mathbf{m})\right\} \qquad (4\text{-}20)$$

where

$$\mathbf{m} = \left[\begin{array}{c} \mathbf{m}_x(t_1) \\ \hline \mathbf{m}_x(t_2) \end{array}\right] = \left[\begin{array}{c} E[\mathbf{x}(t_1)] \\ E[\mathbf{x}(t_2)] \end{array}\right] \qquad (4\text{-}21\text{a})$$

$$\mathbf{P} = \left[\begin{array}{c|c} E[\mathbf{x}(t_1)\mathbf{x}^{\mathrm{T}}(t_1)] - \mathbf{m}_x(t_1)\mathbf{m}_x^{\mathrm{T}}(t_1) & E[\mathbf{x}(t_1)\mathbf{x}^{\mathrm{T}}(t_2)] - \mathbf{m}_x(t_1)\mathbf{m}_x^{\mathrm{T}}(t_2) \\ \hline E[\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_1)] - \mathbf{m}_x(t_2)\mathbf{m}_x^{\mathrm{T}}(t_1) & E[\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_2)] - \mathbf{m}_x(t_2)\mathbf{m}_x^{\mathrm{T}}(t_2) \end{array}\right]$$

$$= \left[\begin{array}{c|c} E[\mathbf{x}(t_1)\mathbf{x}^{\mathrm{T}}(t_1)] & E[\mathbf{x}(t_1)\mathbf{x}^{\mathrm{T}}(t_2)] \\ \hline E[\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_1)] & E[\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_2)] \end{array}\right] - \mathbf{m}\mathbf{m}^{\mathrm{T}} \qquad (4\text{-}21\text{b})$$

Analogous statements could then be made about density functions of any order, corresponding to any choice of N time points instead of just two.

4.3 STATIONARY STOCHASTIC PROCESSES AND POWER SPECTRAL DENSITY

One particularly pertinent characterization of a stochastic process is whether or not it is stationary. In this regard, there is a strict sense of stationarity, concerned with all moments, and a wide sense of stationarity, concerned only with the first two moments. A process $\mathbf{x}(\cdot, \cdot)$ is *strictly stationary* if, for all sets $t_1, \ldots, t_N \in T$ and any $\tau \in T$ [supposing that $(t_i + \tau) \in T$ also], the joint distribution of $\mathbf{x}(t_1 + \tau), \ldots, \mathbf{x}(t_N + \tau)$ does not depend on the time shift τ: i.e.,

$$P(\{\omega : \mathbf{x}(t_1 + \tau, \omega) \le \boldsymbol{\xi}_1, \ldots, \mathbf{x}(t_N + \tau, \omega) \le \boldsymbol{\xi}_N\})$$
$$= P(\{\omega : \mathbf{x}(t_1, \omega) \le \boldsymbol{\xi}_1, \ldots, \mathbf{x}(t_N, \omega) \le \boldsymbol{\xi}_N\}) \qquad (4\text{-}22)$$

A process $\mathbf{x}(\cdot,\cdot)$ is *wide-sense stationary* if, for all t, $\tau \in T$, the following three criteria are met:

   (i) $E\{\mathbf{x}(t)\mathbf{x}(t)^T\}$ is finite.

   (ii) $E\{\mathbf{x}(t)\}$ is a constant.

   (iii) $E\{[\mathbf{x}(t) - \mathbf{m}_x][\mathbf{x}(t+\tau) - \mathbf{m}_x]^T\}$ depends only on the time difference τ [and thus $\mathbf{\Psi}_{xx}(t) = \mathbf{\Psi}_{xx}(t,t)$ and $\mathbf{P}_{xx}(t) = \mathbf{P}_{xx}(t,t)$ are constant].

Thus, in general, a strictly stationary process is wide-sense stationary if and only if it has finite second moments, and wide-sense stationarity does not imply strict-sense stationarity. However, in the special case of Gaussian processes, a wide-sense stationary process is strict-sense stationary as well.

By definition, the correlation and covariance kernels $\mathbf{\Psi}_{xx}(t, t+\tau)$ and $\mathbf{P}_{xx}(t, t+\tau)$ for wide-sense stationary $\mathbf{x}(\cdot,\cdot)$ are functions only of the time difference τ; this is often made explicit notationally by writing these as functions of a single argument, the time difference τ:

$$\mathbf{\Psi}_{xx}(t, t+\tau) \to \mathbf{\Psi}_{xx}(\tau) \tag{4-23a}$$

$$\mathbf{P}_{xx}(t, t+\tau) \to \mathbf{P}_{xx}(\tau) \tag{4-23b}$$

To avoid confusion between these relations and $\mathbf{\Psi}_{xx}(t)$ and $\mathbf{P}_{xx}(t)$ as defined in (4-7), a single argument τ will be reserved to denote a time *difference* when discussing functions $\mathbf{\Psi}_{xx}(\cdot)$ or $\mathbf{P}_{xx}(\cdot)$. A further characterization of these functions can be made as well: not only are the diagonal terms of $\mathbf{\Psi}_{xx}$ functions only of τ, but they are *even* functions of τ that assume their maximum value at $\tau = 0$.

In the case of wide-sense stationary processes, Fourier transform theory can be exploited to generate a frequency-domain characterization of processes in the form of power spectral densities. If a scalar time function $y(\cdot)$ is Fourier transformable, the relation between it and its Fourier transform $\bar{y}(\cdot)$, as a function of frequency ω, is given by

$$\bar{y}(\omega) = a \int_{-\infty}^{\infty} y(t)e^{-j\omega t}\, dt \tag{4-24a}$$

$$y(t) = b \int_{-\infty}^{\infty} \bar{y}(\omega)e^{j\omega t}\, d\omega \tag{4-24b}$$

where a and b are scalars such that their product is $1/(2\pi)$:

$$ab = 1/(2\pi) \tag{4-24c}$$

Power spectral density of a scalar wide-sense stationary process $x(\cdot,\cdot)$ is defined as the Fourier transform of the correlation function $\mathbf{\Psi}_{xx}(\tau)$. Since samples from a wide-sense stationary process must be visualized as existing for all negative and positive time if τ is to be allowed to assume any value in R^1, the two-sided Fourier transform, i.e., integrating from $t = -\infty$ to $t = \infty$, does make sense conceptually in the definition.

Some useful properties of Fourier transforms of real-valued functions include the following:

(1) $\bar{y}(\omega)$ is complex in general, with $\bar{y}(-\omega) = \bar{y}^*(\omega)$, where * denotes complex conjugate.

(2)   Thus, $\bar{y}(\omega)\bar{y}(-\omega) = \bar{y}(\omega)\bar{y}^*(\omega) = |\bar{y}(\omega)|^2$.

(3)   The real part of $\bar{y}(\cdot)$ is an even function of ω, and the imaginary part is odd in ω.

(4)   Although $\delta(t)$ is not Fourier transformable, its transform can be defined formally through (4-24a) as

$$\bar{\delta}(\omega) = a \int_{-\infty}^{\infty} \delta(t)e^{-j\omega t}\, dt = a \tag{4-25a}$$

so that (4-24b) yields, formally,

$$\delta(t) = b \int_{-\infty}^{\infty} ae^{j\omega t}\, d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega t}\, d\omega \tag{4-25b}$$

Unfortunately, there are a number of conventions on the choice of a and b in (4-24) for defining power spectral density. The most common convention is

$$\bar{\Psi}_{xx}(\omega) = \int_{-\infty}^{\infty} \Psi_{xx}(\tau)e^{-j\omega\tau}\, d\tau \tag{4-26a}$$

$$\Psi_{xx}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \bar{\Psi}_{xx}(\omega)e^{j\omega\tau}\, d\omega \tag{4-26b}$$

Note that if frequency is expressed in hertz rather than rad/sec, using $f = \omega/(2\pi)$, then there is a unity coefficient for both defining equations:

$$\bar{\Psi}_{xx}(f) = \int_{-\infty}^{\infty} \Psi_{xx}(\tau)e^{-j2\pi f\tau}\, d\tau \tag{4-27a}$$

$$\Psi_{xx}(\tau) = \int_{-\infty}^{\infty} \bar{\Psi}_{xx}(f)e^{j2\pi f\tau}\, df \tag{4-27b}$$

Using this convention, power spectral density is typically specified in units of (quantity)^2/hertz. Since $\Psi_{xx}(0)$ is just the mean squared value of $x(t)$, it can be obtained by integrating the power spectral density function:

$$E\{x(t)^2\} = \Psi_{xx}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \bar{\Psi}_{xx}(\omega)\, d\omega = \int_{-\infty}^{\infty} \bar{\Psi}_{xx}(f)\, df \tag{4-28}$$

Furthermore, we can use Euler's identity to write (4-26a) as

$$\bar{\Psi}_{xx}(\omega) = \int_{-\infty}^{\infty} \Psi_{xx}(\tau)[\cos \omega\tau - j\sin \omega\tau]\, d\tau$$

and, since $\Psi_{xx}(\tau)$ and $\cos \omega\tau$ are even functions of τ and $\sin \omega\tau$ is odd in τ, this becomes

$$\bar{\Psi}_{xx}(\omega) = \int_{-\infty}^{\infty} \Psi_{xx}(\tau) \cos \omega\tau \, d\tau \tag{4-29a}$$

$$= 2 \int_{0}^{\infty} \Psi_{xx}(\tau) \cos \omega\tau \, d\tau \qquad [\text{if } \Psi_{xx}(0) \text{ is finite}] \tag{4-29b}$$

Thus, the power spectral density function is a *real, even* function of ω. It can be shown to be a pointwise *positive* function of ω as well. Analogously to (4-29), the correlation function becomes

$$\Psi_{xx}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \bar{\Psi}_{xx}(\omega) \cos \omega\tau \, d\omega \tag{4-30a}$$

$$= \frac{1}{\pi} \int_{0}^{\infty} \bar{\Psi}_{xx}(\omega) \cos \omega\tau \, d\omega \qquad [\text{if } \bar{\Psi}_{xx}(0) \text{ is finite}] \tag{4-30b}$$

Another common convention for power spectral density is

$$\bar{\Psi}'_{xx}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_{xx}(\tau) e^{-j\omega\tau} \, d\tau \tag{4-31a}$$

$$\Psi_{xx}(\tau) = \int_{-\infty}^{\infty} \bar{\Psi}'_{xx}(\omega) e^{j\omega\tau} \, d\omega \tag{4-31b}$$

with units as (quantity)^2/(rad/sec), motivated by the property that the mean squared value becomes

$$E\{x(t)^2\} = \int_{-\infty}^{\infty} \bar{\Psi}'_{xx}(\omega) \, d\omega \tag{4-32}$$

i.e., without the $1/(2\pi)$ factor. A third convention encountered commonly is

$$\bar{\Psi}''_{xx}(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \Psi_{xx}(\tau) e^{-j\omega\tau} \, d\tau \tag{4-33a}$$

$$\Psi_{xx}(\tau) = \frac{1}{2} \int_{-\infty}^{\infty} \bar{\Psi}''_{xx}(\omega) e^{j\omega\tau} \, d\omega \tag{4-33b}$$

also with units as (quantity)^2/(rad/sec), such that

$$E\{x(t)^2\} = \int_{0}^{\infty} \bar{\Psi}''_{xx}(\omega) \, d\omega \tag{4-34}$$

using only *positive* values of ω to correspond to a physical interpretation of frequencies being nonnegative.

The name power spectral density can be motivated by interpreting "power" in the generalized sense of expected squared values of the members of an ensemble. $\bar{\Psi}_{xx}(\omega)$ is a spectral density for the power in a process $x(\cdot)$ in that integration of $\bar{\Psi}_{xx}$ over the frequency band from ω_1 to ω_2 yields the mean squared value of the process which consists only of those harmonic com-
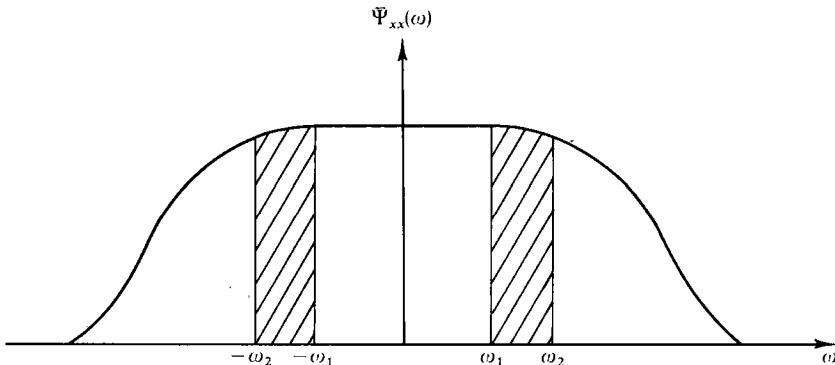
FIG. 4.5   Power spectral density.

ponents of $x(t)$ that lie between ω_1 and ω_2, as shown by the shaded region in Fig. 4.5. (This is in fact another means of defining power spectral density.) In particular, the mean squared value of $x(t)$ itself is given by an integration of $\Psi_{xx}(\omega)$ over the full range of possible frequencies ω.

EXAMPLE 4.2   Figure 4.6 depicts the autocorrelation functions and power spectral density functions (using the most common convention of definition) of a white process, an exponentially time-correlated process, and a random bias. Note that a white noise is uncorrelated in time, yielding an impulse at $\tau = 0$ in Fig. 4.6a; the corresponding power spectral density is flat over all ω—equal power content over all frequencies. Figure 4.6b corresponds to an exponentially time-correlated process with correlation time T, as discussed in Example 4.1. Heuristically, these
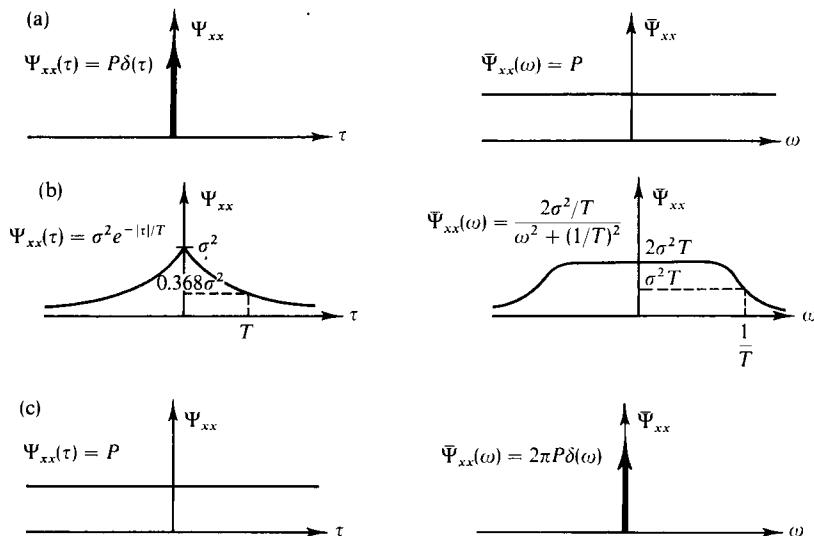
FIG. 4.6   Typical autocorrelations and power spectral densities. (a) White process. (b) Exponentially time-correlated process. (c) Random bias.

plots converge to those of (a) as $T \to 0$, and to those of (c) as $T \to \infty$. Figure 4.6c corresponds to a random bias process, the samples of which are constant in time—thus, there is constant correlation over all time differences τ, and all of the process power is concentrated at the zero frequency. Another process with nonzero power at a *discrete* frequency would be the process composed of sinusoids at a known frequency ω_0 and of uniformly distributed phase, with power spectral density composed of two impulses, at ω_0 and $-\omega_0$, and a cosinusoidal autocorrelation.   ∎

The *cross-power spectral density* of two wide-sense stationary scalar processes $x(\cdot, \cdot)$ and $y(\cdot, \cdot)$ is the Fourier transform of the associated cross-correlation function:

$$\Psi_{xy}(\omega) = \int_{-\infty}^{\infty} \Psi_{xy}(\tau) e^{-j\omega\tau}\, d\tau \tag{4-35}$$

This is, in general, a complex function of ω, and, since $\Psi_{xy}(\tau) = \Psi_{yx}(-\tau)$, the following relations are valid.

$$\Psi_{xy}(\omega) = \Psi_{yx}^{*}(\omega) \tag{4-36}$$

$$\tfrac{1}{2}[\Psi_{xy}(\omega) + \Psi_{yx}(\omega)] = \mathrm{Real}\{\Psi_{xy}(\omega)\} = \mathrm{Real}\{\Psi_{yx}(\omega)\} \tag{4-37}$$

One subclass of real strictly stationary processes of particular interest is the set of ergodic processes. A process is *ergodic* if any statistic calculated by averaging over all members of the ensemble of samples at a fixed time can be calculated equivalently by time-averaging over any single representative member of the ensemble, except possibly a single member out of a set of probability zero. Not all stationary processes are ergodic: the ensemble of constant functions is an obvious counterexample, in that a time average of each sample will yield that particular constant value rather than the mean value for the entire ensemble.

There is no readily applied condition to ensure ergodicity in general. For a scalar stationary Gaussian process $x(\cdot, \cdot)$ defined on $T = (-\infty, \infty)$, a sufficient condition does exist: $x(\cdot, \cdot)$ is ergodic if $\int_{-\infty}^{\infty} |P_{xx}(\tau)|\, d\tau$ is finite [however, $P_{xx}(\tau)$ itself requires an ensemble average]. In practice, empirical results for stationary processes are often obtained by time-averaging of a single process sample, under the *assumption* of ergodicity, such as

$$m_x = E[x(t, \cdot)] = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t, \omega_i)\, dt \tag{4-38a}$$

$$\Psi_{xx}(\tau) \triangleq E[x(t, \cdot)x(t + \tau, \cdot)]$$

$$= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t, \omega_i)x(t + \tau, \omega_i)\, dt \tag{4-38b}$$

$$\Psi_{xy}(\tau) = E[x(t, \cdot)y(t + \tau, \cdot)]$$

$$= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t, \omega_i)y(t + \tau, \omega_i)\, dt \tag{4-38c}$$

Moreover, these are only approximately evaluated due to the use of finite length, rather than infinite, samples.

4.4   SYSTEM MODELING: OBJECTIVES AND DIRECTIONS

Suppose we are given a physical system that can be subjected to known controls and to inputs beyond our own direct control, typically wideband noises (noises with instantaneous power over a wide range of frequencies), although narrow band noises and other forms are also possible. Further assume that we want to characterize certain outputs of the system, for instance, by depicting their mean and covariance kernel for all time values. Such a characterization would be necessary to initiate designs of estimators or controllers for the system, and a prerequisite to a means of analyzing the performance capabilities of such devices as well.

The objective of a mathematical model would be to generate an adequate, tractable representation of the behavior of all outputs of interest from the real physical system. Here "adequate" is subjective and is a function of the intended use of this representation. For example, if a gyro were being tested in a laboratory, one would like to develop a mathematical model that would generate outputs whose characteristics were identical to those actually observed empirically. Since no model is perfect, one really attempts to generate models that closely approximate the behavior of observed quantities.

From deterministic modeling, we gain the insight that a potentially useful model form would be a linear state equation and sampled data output relation *formally* written as

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{n}_1(t) \tag{4-39a}$$

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{n}_2(t_i) \tag{4-39b}$$

These are direct extensions of Eqs. (2-35) and (2-60) obtained by adding a noise process $\mathbf{n}_1(\cdot,\cdot)$ to the dynamics equation and $\mathbf{n}_2(\cdot,\cdot)$ to the output equation, with $\mathbf{n}_1(t,\omega_k) \in R^s$ and $\mathbf{n}_2(t_i,\omega_j) \in R^m$. [$\mathbf{G}(t)$ is n-by-s to be compatible with the state dimension n, and $\mathbf{n}_2(t_i,\omega_j)$ is of dimension m, the number of measurements available.] Note that (4-39b) could be extended to allow direct feedthrough of $\mathbf{u}$ and $\mathbf{n}_1$, but this will not be pursued here. Again it is emphasized that (4-39) is just a formal extension; for instance, how would $\dot{\mathbf{x}}(t)$ be interpreted fundamentally?

Unfortunately, a model of this generality is not directly exploitable. We would like to evaluate the joint probability distribution or density function for $\mathbf{x}(t_1,\cdot),\ldots,\mathbf{x}(t_N,\cdot)$, and, through this evaluation, to obtain the corresponding joint functions for $\mathbf{z}(t_1,\cdot),\ldots,\mathbf{z}(t_N,\cdot)$. This is generally infeasible. For example, if $\mathbf{n}_1(t)$ were uniformly distributed for all $t \in T$, one can say very little

about these joint probability functions. However, if $\mathbf{n}_1(\cdot,\cdot)$ and $\mathbf{n}_2(\cdot,\cdot)$ were assumed *Gaussian*, then all distribution or density functions of interest might be shown to be Gaussian as well, completely characterized by the corresponding first two moments. If observed quantities are not Gaussian, one can still seek to construct a model that provides a Gaussian output whose first two moments duplicate the first two moments of the empirically observed data.

Complete depiction of a joint distribution or density is still generally intractable however. In order to achieve this objective, we will restrict attention to system inputs describable as *Markov processes*. Let $\mathbf{x}(\cdot,\cdot)$ be a random process and consider

$$F_{\mathbf{x}(t_i)|\mathbf{x}(t_{i-1}),\,\mathbf{x}(t_{i-2}),\,\ldots,\,\mathbf{x}(t_j)}(\boldsymbol{\xi}_i\,|\,\mathbf{x}_{i-1},\mathbf{x}_{i-2},\ldots,\mathbf{x}_j)$$

i.e., the probability distribution function of $\mathbf{x}(t_i)$ as a function of the n-vector $\boldsymbol{\xi}_i$, given that $\mathbf{x}(t_{i-1},\omega_k) = \mathbf{x}_{i-1}$, $\mathbf{x}(t_{i-2},\omega_k) = \mathbf{x}_{i-2}, \ldots, \mathbf{x}(t_j,\omega_k) = \mathbf{x}_j$. If

$$\begin{aligned}
F_{\mathbf{x}(t_i)|\mathbf{x}(t_{i-1}),\,\mathbf{x}(t_{i-2}),\,\ldots,\,\mathbf{x}(t_j)}&(\boldsymbol{\xi}_i\,|\,\mathbf{x}_{i-1},\mathbf{x}_{i-2},\ldots,\mathbf{x}_j) \\
&= F_{\mathbf{x}(t_i)|\mathbf{x}(t_{i-1})}(\boldsymbol{\xi}_i\,|\,\mathbf{x}_{i-1})
\end{aligned} \tag{4-40}$$

for any countable choice of values i and j and for all values of $\mathbf{x}_{i-1}, \ldots, \mathbf{x}_j$, then $\mathbf{x}(\cdot,\cdot)$ is a Markov process. Thus, the Markov property for stochastic processes is conceptually analogous to the ability to define a system *state* for deterministic processes. The value that the process $\mathbf{x}(\cdot,\cdot)$ assumes at time t_{i-1} provides as much information about $\mathbf{x}(t_i,\cdot)$ as do the values of $\mathbf{x}(\cdot,\cdot)$ at time t_{i-1} and all previous time instants: the value assumed by $\mathbf{x}(t_{i-1},\cdot)$ embodies all information needed for propagation to time t_i, and the past history leading to $\mathbf{x}_{i-1}$ is of no consequence. In the context of linear system models, the Markov assumption will be shown equivalent to the fact that the continuous-time process $\mathbf{n}_1(\cdot,\cdot)$ and the discrete-time process $\mathbf{n}_2(\cdot,\cdot)$ in (4-39) are expressible as the outputs of linear state-described models, called "shaping filters," driven only by deterministic inputs and *white noises*. A *Gauss–Markov process* is then a process which is both Gaussian and Markov.

Thus, the form of the system model depicted in Fig. 4.7 is motivated. A linear model of the physical system is driven by deterministic inputs, white Gaussian noises, and Gauss–Markov processes. As discussed in Section 1.4, the white noises are chosen as adequate representations of wideband noises with essentially constant power density over the system bandpass. The other Markov processes are time-correlated processes for which a white model would be inadequate. However, these can be generated by passing white noise through linear shaping filters. Consequently, one can consider the original system model and the shaping filters as a single "augmented" linear system, driven *only* by deterministic inputs and white Gaussian noises. This can be described through a restricted form of (4-39):

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \tag{4-41a}$$

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{4-41b}$$

Augmented system model (linear)
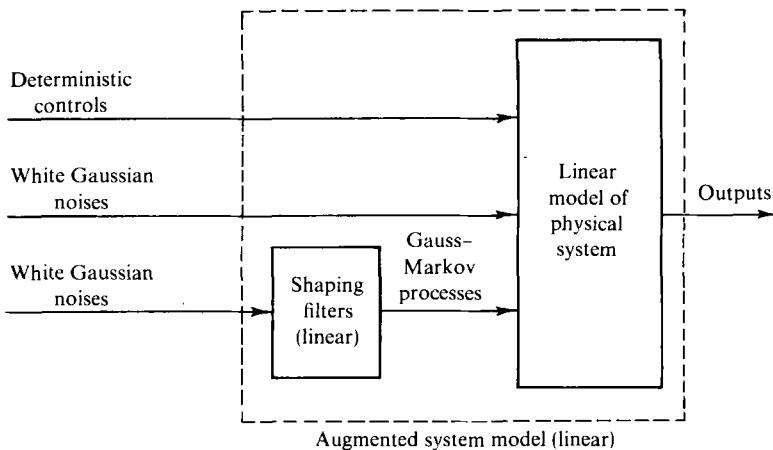
FIG. 4.7   Linear system model.

where $\mathbf{x}(t)$ is now the augmented system state, and $\mathbf{w}(t)$ and $\mathbf{v}(t_i)$ are white Gaussian noises, assumed independent of each other and of the initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$, where $\mathbf{x}_0$ is a Gaussian random variable. These noises model not only the disturbances and noise corruption that affect the system, but also the uncertainty inherent in the mathematical models themselves.

Using the insights from deterministic system theory, one seeks a solution to (4-41a): formally, one could write

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau + \int_{t_0}^{t} \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{w}(\tau)\,d\tau \quad (4\text{-}42)$$

If this were a valid result, then it would be possible to describe the first two moments of $\mathbf{x}(\cdot, \cdot)$ and thus totally characterize this Gaussian stochastic process. However, the last term in (4-42) cannot be evaluated properly, and thus (4-42) has no real meaning at all. The remainder of this chapter is devoted to (1) a proper development of the mathematics as motivated by this section and (2) the practical application of the results to useful model formulations.

4.5   FOUNDATIONS: WHITE GAUSSIAN NOISE AND BROWNIAN MOTION

The previous section motivated the use of white Gaussian noise models as the only stochastic inputs to a linear system model, and this warrants further attention. First, a process $\mathbf{x}(\cdot, \cdot)$ is a *white Gaussian process* if, for any choice of $t_1, \ldots, t_N \in T$, the N random vectors $\mathbf{x}(t_1, \cdot), \ldots, \mathbf{x}(t_N, \cdot)$ are independent Gaussian random vectors. If the time set of interest, T, is a set of discrete time points, this is conceptually straightforward and implies that

$$\mathbf{P}_{xx}(t_i, t_j) = \mathbf{0} \qquad \text{if} \quad i \neq j \tag{4-43}$$

Such a discrete-time process can in fact be used to drive a difference equation model of a system with no theoretical difficulties.

However, if T is a time interval, then the definition of a white Gaussian process implies that there is no correlation between $\mathbf{x}(t_i, \cdot)$ and $\mathbf{x}(t_j, \cdot)$, even for t_i and t_j separated by only an infinitesimal amount:

$$E\{\mathbf{x}(t_i)\mathbf{x}^\mathrm{T}(t_j)\} = \mathbf{\Psi}_{xx}(t_i)\delta(t_i - t_j) \tag{4-44}$$

This is contrary to the behavior exhibited by any processes observed empirically. If we consider stationary white Gaussian noise [not wide sense to be precise, since $E\{\mathbf{x}(t_i)\mathbf{x}^\mathrm{T}(t_i)\}$ is not finite], then the power spectral density of such a process would be constant over *all* frequencies, and thus it would be an infinite power process: thus it cannot exist. Moreover, if we were to construct a continuous-time system model in the form of a linear differential equation driven by such a process, then a solution to the differential equation could *not* be obtained rigorously, as pointed out in the last section.

Brownian motion (or the "Wiener process") [8, 13] will serve as a basic process for continuous-time modeling. Through it, system models can be *properly* developed in the form of stochastic differential equations whose solutions *can* be obtained. Scalar constant-diffusion Brownian motion will be discussed first, and then extensions made to the general case of a vector time-varying-diffusion Brownian motion process.

To discuss Brownian motion, we first need a definition of a *process with independent increments*. Let $t_0 < t_1 < \cdots < t_N$ be a partition of the time interval T. If the "increments" of the process $\mathbf{x}(\cdot, \cdot)$, i.e., the set of N random variables

$$\begin{aligned}
\boldsymbol{\delta}_1(\cdot) &= [\mathbf{x}(t_1, \cdot) - \mathbf{x}(t_0, \cdot)] \\
\boldsymbol{\delta}_2(\cdot) &= [\mathbf{x}(t_2, \cdot) - \mathbf{x}(t_1, \cdot)] \\
&\;\;\vdots \\
\boldsymbol{\delta}_N(\cdot) &= [\mathbf{x}(t_N, \cdot) - \mathbf{x}(t_{N-1}, \cdot)]
\end{aligned} \tag{4-45}$$

are mutually independent for *any* such partition of T, then $\mathbf{x}(\cdot, \cdot)$ is said to be a process with independent increments.

The process $\beta(\cdot, \cdot)$ is defined to be a *scalar constant-diffusion Brownian motion process* if

(i)   it is a process with independent increments,

(ii)   the increments are Gaussian random variables such that, for t_1 and t_2 any time instants in T,

$$E\{[\beta(t_2) - \beta(t_1)]\} = 0 \tag{4-46a}$$

$$E\{[\beta(t_2) - \beta(t_1)]^2\} = q|t_2 - t_1| \tag{4-46b}$$

(iii) $\beta(t_0, \omega_i) = 0$ for all $\omega_i \in \Omega$, except possibly a set of ω_i of probability zero (this is by *convention*).

Such a definition provides a mathematical abstraction of empirically observed Brownian motion processes, such as the motion of gas molecules. Figure 4.8 depicts some samples from a Brownian motion process $\beta(\cdot,\cdot)$. Note that for all samples shown, $\beta(t_0,\omega_i) = 0$. Specific realizations of increments are also shown, for instance,

$$\Delta_1 = \delta_1(\omega_2) = [\beta(t_2,\omega_2) - \beta(t_1,\omega_2)]$$
$$\Delta_2 = \delta_2(\omega_2) = [\beta(t_3,\omega_2) = \beta(t_2,\omega_2)]$$

$$(4\text{-}47)$$

The random variables $\delta_1(\cdot)$ and $\delta_2(\cdot)$ are independent.
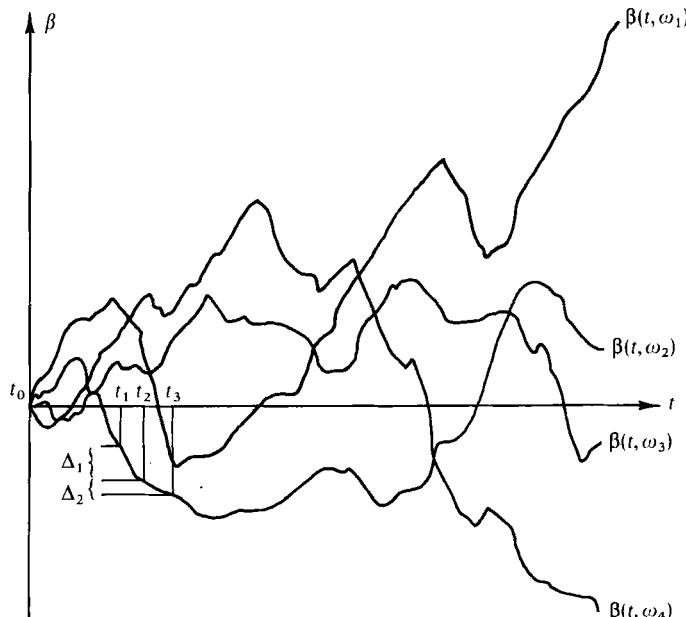
FIG. 4.8   Samples from a Brownian motion process.

The parameter q in (4-46b) is called the *diffusion* of the process. Note that the variance of the change in value of $\beta(\cdot,\cdot)$ between any time t_1 and any later time t_2 is a *linear* function of the *time difference* $(t_2 - t_1)$. For that reason, constant-diffusion Brownian motion is sometimes termed (misleadingly) "stationary" Brownian motion, but such terminology will be avoided here.

Since $\beta(t_1,\cdot)$ for given $t_1 \in T$ is a random variable composed of a sum of independent Gaussian increments, it is also Gaussian, with statistics

$$m_\beta(t_i) = E[\beta(t_i)] = 0 \qquad (4\text{-}48a)$$

$$P_{\beta\beta}(t_i) = E[\beta(t_i)^2] = q[t_i - t_0] \qquad (4\text{-}48b)$$

Thus q indicates how fast the mean square value of $\beta(\cdot,\cdot)$ diverges from its initial value of zero at time t_0.

To characterize the scalar constant-diffusion Brownian motion completely, we can explicitly generate the joint density functions for any finite set of random variables $\beta(t_0,\cdot), \ldots, \beta(t_N,\cdot)$ as a Gaussian density function, with zero mean and covariance composed of diagonal terms as $P_{\beta\beta}(t_i, t_i) = P_{\beta\beta}(t_i)$ as in (4-48b). The off-diagonal terms can be specified by considering $t_j > t_i$ and writing

$$\beta(t_j) = \beta(t_i) + [\beta(t_j) - \beta(t_i)]$$

so that the off-diagonal elements become

$$\begin{aligned} E\{\beta(t_i)\beta(t_j)\} &= E\{\beta(t_i)^2\} + E\{\beta(t_i)[\beta(t_j) - \beta(t_i)]\} \\ &= E\{\beta(t_i)^2\} + E\{\beta(t_i)\}E\{[\beta(t_j) - \beta(t_i)]\} \\ &= E\{\beta(t_i)^2\} = q(t_i - t_0) \end{aligned} \qquad (4\text{-}49)$$

where the second equality follows from independence of $\beta(t_i)$ and $[\beta(t_j) - \beta(t_i)]$ for $t_j > t_i$, and then both separate expectations are zero.

Although scalar constant-diffusion Brownian motion was just described completely in a probabilistic sense, a further characterization in terms of such concepts as continuity and differentiability is desirable, since these will directly influence the development and meaning of stochastic differential equations. For a deterministic function f, such concepts are straightforwardly approached by asking if the number $f(t_2)$ converges to the number $f(t_1)$ in the limit as t_2 approaches t_1, or similarly if an appropriate difference quotient converges to some limit. For stochastic processes, one needs to conceive of what "convergence" itself means. There are three *concepts of convergence* [4,9,13] of use to us: (1) mean square convergence, (2) convergence in probability, and (3) convergence almost surely (with probability one).

A sequence of random variables, $x_1, x_2, \ldots$, is said to *converge in mean square* (or sometimes, to converge "in the mean") to the random variable x if $E[x_k^2]$ is finite for all k, and $E[x^2]$ is finite, and

$$\lim_{k \to \infty} E[(x_k - x)^2] = 0 \qquad (4\text{-}50)$$

Thus, we are again concerned with the convergence of a sequence of *real numbers* $E[(x_k - x)^2]$ in order to establish convergence in mean square. If (4-50) holds, then one often writes

$$\underset{k \to \infty}{\text{l.i.m.}}\ x_k = x \qquad (4\text{-}51)$$

where l.i.m. denotes limit in the mean. This conception of convergence will provide the basis for defining stochastic integrals subsequently.

A sequence of random variables $x_1, x_2, \ldots$ is said to *converge in probability* to x if, for all $\varepsilon > 0$,

$$\lim_{k \to \infty} P(\{\omega : |x_k(\omega) - x(\omega)| \geq \varepsilon\}) = 0 \qquad (4\text{-}52)$$

Here the sequence of real numbers is a sequence of individual probabilities, which is to converge to zero no matter how small ε might be chosen.

A sequence of random variables $x_1, x_2, \ldots$ is said to *converge almost surely* (a.s.) or to *converge with probability one* (w.p.1) to x if

$$\lim_{k \to \infty} |x_k(\omega) - x(\omega)| = 0 \qquad (4\text{-}53)$$

for "almost all" realizations: if (4-53) holds for all ω except possibly a set A of ω whose probability is zero, $P(A) = 0$. Unlike the two previous concepts, this idea of convergence directly considers the convergence of every sequence of *realizations* of the random variables involved, rather than ensemble averages or probabilities.

Convergence in probability is the weakest of the three concepts, and it can be shown to be implied by the others:

$$[\text{Convergence in mean square}] \rightarrow [\text{Convergence in probability}] \quad (4\text{-}54\text{a})$$

$$[\text{Convergence almost surely}] \rightarrow [\text{Convergence in probability}] \quad (4\text{-}54\text{b})$$

Relation (4-54a) follows directly from the *Chebychev inequality*:

$$P(\{\omega : |x_k(\omega) - x(\omega)| \geq \varepsilon\}) \leq E\{[x_k(\cdot) - x(\cdot)]^2\}/\varepsilon^2 \qquad (\text{all } \varepsilon > 0) \quad (4\text{-}55)$$

since, if the mean square limit exists, then $\lim_{k \to \infty} E\{[x_k - x]^2\} = 0$, and so $\lim_{k \to \infty} P(\{\omega : |x_k(\omega) - x(\omega)| \geq \varepsilon\}) = 0$ for all $\varepsilon > 0$. Convergence in mean square does not imply, and is not implied by, convergence almost surely.

The *continuity of Brownian motion* can now be described. (For proof, see [4].) Let $\beta(\cdot, \cdot)$ be a Brownian motion process defined on $T \times \Omega$ with $T = [0, \infty)$. Then to each point $t \in T$ there corresponds two random variables $\beta^-(t, \cdot)$ and $\beta^+(t, \cdot)$ such that

$$\underset{t' \uparrow t}{\text{l.i.m.}} \; \beta(t', \cdot) = \beta^-(t, \cdot) \qquad (4\text{-}56\text{a})$$

$$\underset{t' \downarrow t}{\text{l.i.m.}} \; \beta(t', \cdot) = \beta^+(t, \cdot) \qquad (4\text{-}56\text{b})$$

where $t' \uparrow t$ means in the limit as t' approaches t from below and $t' \downarrow t$ means as t' approaches t from above. Furthermore, for each $t \in T$,

$$\beta^-(t, \cdot) = \beta(t, \cdot) = \beta^+(t, \cdot) \qquad (4\text{-}56\text{c})$$

almost surely. Equation (4-56b) states that as we let time t' approach time t from above, the value of $E\{[\beta(t', \cdot) - \beta^+(t, \cdot)]^2\}$ converges to zero: the variance

describing the spread of values of realizations of $\beta(t', \cdot)$ from realizations of $\beta^+(t, \cdot)$ goes to zero in the limit. Moreover, (4-56c) dictates that all realizations $\beta^+(t, \omega_i)$ equal $\beta(t, \omega_i)$, except possibly for a set of ω_i's whose total probability is zero. Similar results are obtained by letting t' approach t from below as well.

This result implies that $\beta(\cdot, \cdot)$ is also continuous in probability through the Chebychev inequality:

$$P(\{\omega : |\beta(t', \cdot) - \beta(t, \cdot)| \geq \varepsilon\}) \leq E\{[\beta(t', \cdot) - \beta(t, \cdot)]^2\}/\varepsilon^2 = q|t' - t|/\varepsilon^2 \quad (4\text{-}57)$$

for all $\varepsilon > 0$. Consequently, the limit of the probability in (4-57) is zero as t' approaches t from above or below.

Moreover, Brownian motion can be shown to be continuous almost surely. In other words, almost all samples from the process (except possibly a set of samples of probability zero) are themselves continuous.

Brownian motion is nondifferentiable in the mean square and almost sure senses. A process $x(\cdot, \cdot)$ is mean square differentiable if the limit

$$\underset{\Delta t \to 0}{\text{l.i.m.}} \frac{x(t + \Delta t, \cdot) - x(t, \cdot)}{\Delta t}$$

exists, and then this limit defines the mean square derivative, $\dot{x}(t, \cdot)$, at $t \in T$. However, for Brownian motion $\beta(\cdot, \cdot)$,

$$E\left\{\left[\frac{\beta(t + \Delta t, \cdot) - \beta(t, \cdot)}{\Delta t}\right]\right\} = 0 \quad (4\text{-}58\text{a})$$

$$E\left\{\left[\frac{\beta(t + \Delta t, \cdot) - \beta(t, \cdot)}{\Delta t}\right]^2\right\} = \frac{q\,\Delta t}{\Delta t^2} = \frac{q}{\Delta t} \quad (4\text{-}58\text{b})$$

Thus, as $\Delta t \to 0$, the variance of the difference quotient used to define the mean square derivative becomes infinite. This can be used to show that

$$\lim_{\Delta t \to 0} P\left(\left\{\omega : \left|\frac{\beta(t + \Delta t, \omega) - \beta(t, \omega)}{\Delta t}\right| \leq B\right\}\right) = 0 \quad (4\text{-}59)$$

for any finite choice of bound B. Thus, the difference quotient for defining the derivative of each sample function has no finite limit for any $t \in T$ and almost all $\omega \in \Omega$ (except possibly for a set of probability zero): Brownian motion is nondifferentiable almost surely.

Thus, almost all sample functions from a Brownian motion process are continuous but nondifferentiable. Heuristically they are continuous, but have "corners" everywhere. Moreover, it can be shown that these sample functions are also of unbounded variation with probability one. It is this property especially that precludes a fruitful development of stochastic integrals in an almost sure sense. Instead, we will pursue a mean square approach to stochastic integral and differential equations.

Having described scalar constant-diffusion Brownian motion, it is now possible to investigate continuous-time *scalar stationary white Gaussian noise*. Let us assume (incorrectly) that the Brownian motion process $\beta(\cdot, \cdot)$ is differentiable, and that there is an integrable process $w(\cdot, \cdot)$ such that, for $t, \tau \in T$,

$$\beta(t, \cdot) = \int_{t_0}^{t} w(\tau, \cdot) d\tau \tag{4-60}$$

where the integral is to be understood in some as yet unspecified sense. In other words, we assume that $w(\cdot, \cdot)$ is the derivative of Brownian motion,

$$w(t, \cdot) = d\beta(t, \cdot)/dt \tag{4-61}$$

a derivative that does not really exist. Formal procedures based on this incorrect assumption will reveal that $w(\cdot, \cdot)$ is, in fact, white Gaussian noise.

Let us calculate the mean and variance kernel for this fictitious process. First consider two disjoint time intervals, $(t_1, t_2]$ and $(t_3, t_4]$, so that, formally,

$$\beta(t_2) - \beta(t_1) = \int_{t_1}^{t_2} w(t) dt \tag{4-62a}$$

$$\beta(t_4) - \beta(t_3) = \int_{t_3}^{t_4} w(t) dt \tag{4-62b}$$

Use the properties of Brownian motion and (4-62a) to write

$$E\left\{ \int_{t_1}^{t_2} w(t) dt \right\} = E\{ [\beta(t_2) - \beta(t_1)] \} = 0$$

If the preceding formal integral has the properties of regular integrals, then the expectation operation can be brought inside the time integrals, to yield

$$\int_{t_1}^{t_2} E\{w(t)\} dt = 0$$

Since t_1 and t_2 can be arbitrary, this could only be true if

$$E\{w(t)\} = 0 \qquad \text{for all} \quad t \in T \tag{4-63}$$

To establish the variance kernel, consider the two disjoint intervals, and use the property of independent increments of Brownian motion to write:

$$E\{ [\beta(t_4) - \beta(t_3)][\beta(t_2) - \beta(t_1)] \} = 0$$

Using (4-62), this yields, formally

$$0 = E\left\{ \int_{t_3}^{t_4} w(t) dt \int_{t_1}^{t_2} w(t') dt' \right\}$$

$$= E\left\{ \int_{t_3}^{t_4} \int_{t_1}^{t_2} w(t)w(t') dt' dt \right\}$$

$$= \int_{t_3}^{t_4} \int_{t_1}^{t_2} E\{w(t)w(t')\} dt' dt$$

Since $(t_1, t_2]$ and $(t_3, t_4]$ are arbitrary disjoint intervals, this implies

$$E\{\mathbf{w}(t)\mathbf{w}(t')\} = 0 \qquad \text{for} \quad t \neq t' \tag{4-64}$$

To establish $E\{\mathbf{w}(t)^2\}$, perform the same steps, but using a single interval. By the fact that $\beta(t)$ is Brownian motion,

$$E\{[\beta(t_2) - \beta(t_1)]^2\} = q[t_2 - t_1] = \int_{t_1}^{t_2} q \, dt$$

Combining this and (4-62a) yields

$$\int_{t_1}^{t_2} q \, dt = E\left\{ \int_{t_1}^{t_2} \mathbf{w}(t) \, dt \int_{t_1}^{t_2} \mathbf{w}(t') \, dt' \right\}$$

$$= \int_{t_1}^{t_2} \int_{t_1}^{t_2} E\{\mathbf{w}(t)\mathbf{w}(t')\} \, dt' \, dt$$

or, rewriting,

$$\int_{t_1}^{t_2} \left[\int_{t_1}^{t_2} E\{\mathbf{w}(t)\mathbf{w}(t')\} \, dt' - q \right] dt = 0$$

Since this is true for an arbitrary interval $(t_1, t_2]$, this implies

$$\int_{t_1}^{t_2} E\{\mathbf{w}(t)\mathbf{w}(t')\} \, dt' = q \tag{4-65}$$

for $t \in (t_1, t_2]$.

Now (4-64) and (4-65) together yield the definition of a delta function, so that we can write (4-63)–(4-65) as

$$E\{\mathbf{w}(t)\} = 0 \tag{4-66a}$$

$$E\{\mathbf{w}(t)\mathbf{w}(t')\} = q \, \delta(t - t') \tag{4-66b}$$

Furthermore, $\mathbf{w}(\cdot, \cdot)$ can be shown to be Gaussian, and thus, is a zero-mean, white Gaussian noise process of *strength q*. Heuristically, one can generate Brownian motion of diffusion q by passing white Gaussian noise of strength q through an integrator, as depicted in Fig. 4.9.

The preceding discussion can be generalized to the case of *scalar time-varying-diffusion Brownian motion* by redefining

$$E\{[\beta(t_2) - \beta(t_1)]^2\} = \int_{t_1}^{t_2} q(t) \, dt \tag{4-67}$$
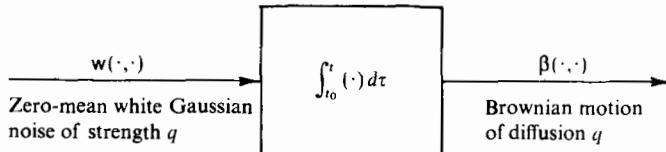
FIG. 4.9   White Gaussian noise and Brownian motion.

for $t_2 \geq t_1$ and $q(t) \geq 0$ for all $t \in T$, instead of $q[t_2 - t_1]$ as in (4-46b). If we assume $q(\cdot)$ to be at least piecewise continuous, which is very nonrestrictive, then no problems are encountered in making the extension. The corresponding *scalar nonstationary white Gaussian noise* $w(\cdot, \cdot)$ would be described by the statistics

$$E\{w(t)\} = 0 \tag{4-68a}$$

$$E\{w(t)w(t')\} = q(t)\, \delta(t - t') \tag{4-68b}$$

for all $t, t' \in T$. This extension will be essential to an adequate description of a "wideband" noise process whose strength can vary with time, typical of many problems of interest. [Note that frequency domain concepts such as a frequency bandwidth cannot be handled rigorously for nonstationary processes, but in many applications $q(\cdot)$ varies slowly with time, and quasi-static methods can be applied.]

EXAMPLE 4.3 The accuracy of position data available to an aircraft from radio navigation aids such as radar, VOR/DME, or TACAN, varies with the range from the aircraft to the navigation aid station. This range data is corrupted by wideband noise, and a reasonable model for indicated range $r_{\text{indicated}}$ is a stochastic process model defined for $t \in T$ and $\omega \in \Omega$ through

$$r_{\text{indicated}}(t, \omega) = r_{\text{true}}(t) + w(t, \omega)$$

where $w(\cdot, \cdot)$ is zero-mean white Gaussian noise, with $q(\cdot)$ reaching a minimum when the aircraft is at minimum distance from the station. The $q(\cdot)$ function of time can be established by knowing the nominal flight path as a function of time. ■

Vector Brownian motion is a further extension, defined as an n-vector stochastic process, $\boldsymbol{\beta}(\cdot, \cdot)$, that has independent Gaussian increments with:

$$E\{\boldsymbol{\beta}(t)\} = \mathbf{0} \tag{4-69a}$$

$$E\{[\boldsymbol{\beta}(t_2) - \boldsymbol{\beta}(t_1)][\boldsymbol{\beta}(t_2) - \boldsymbol{\beta}(t_1)]^{\mathrm{T}}\} = \int_{t_1}^{t_2} \mathbf{Q}(t)\, dt \tag{4-69b}$$

for $t_2 \geq t_1$, and $\mathbf{Q}(t)$ is symmetric and positive semidefinite for all $t \in T$ and $\mathbf{Q}(\cdot)$ is at least piecewise continuous. The corresponding *vector white Gaussian noise* would be the hypothetical time derivative of this vector Brownian motion: a Gaussian process $w(\cdot, \cdot)$ with

$$E\{\mathbf{w}(t)\} = \mathbf{0} \tag{4-70a}$$

$$E\{\mathbf{w}(t)\mathbf{w}^{\mathrm{T}}(t')\} = \mathbf{Q}(t)\delta(t - t') \tag{4-70b}$$

for all $t, t' \in T$, with the same description of $\mathbf{Q}(\cdot)$. Note that (4-70b) indicates that $\mathbf{w}(\cdot, \cdot)$ is uncorrelated in time, which implies that $\mathbf{w}(\cdot, \cdot)$ is white (independent in time) because it is Gaussian. However, this does *not* mean to say that the components of $\mathbf{w}(\cdot, \cdot)$ are uncorrelated with each other at the same time instant: $\mathbf{Q}(t)$ *can* have nonzero off-diagonal terms.

EXAMPLE 4.4   The radio navigation aids described in Example 4.3 provide bearing information as well as range. If b denotes bearing, the data available to the aircraft at time t can be modeled as

$$r_{\text{indicated}}(t, \omega) = r_{\text{true}}(t) + w_1(t, \omega)$$

$$b_{\text{indicated}}(t, \omega) = b_{\text{true}}(t) + w_2(t, \omega)$$

with $w(\cdot, \cdot)$ a zero-mean white Gaussian noise to model the actual wideband noise corruption. The 2-by-2 matrix $Q(t)$ is composed of variances $\sigma_{w_1}^2(t)$ and $\sigma_{w_2}^2(t)$ along the diagonal, with an off-diagonal term of $E\{w_1(t)w_2(t)\}$, generally nonzero.   ∎

4.6   STOCHASTIC INTEGRALS

In Section 4.4, a formal approach to stochastic differential equations led to a solution form (4-42) involving an integral $\int_{t_0}^t \Phi(t, \tau)G(\tau)w(\tau)\,d\tau$ with $w(\cdot, \cdot)$ white Gaussian noise, to which no meaning could be attributed rigorously. From Section 4.5, especially (4-61), one perceives that it may however be possible to give meaning to $\int_{t_0}^t \Phi(t, \tau)G(\tau)\,d\beta(\tau)$ in some manner, thereby generating proper solutions to stochastic differential equations. Consequently, we have to consider the basics of defining integrals as the limit of sums, being careful to establish the conditions under which such a limit in fact exists. To do this properly will require certain concepts from functional analysis, which will be introduced heuristically rather than rigorously. First the simple scalar case is developed in detail, then the general case can be understood as an extension of the same basic concepts.

If $a(\cdot)$ is a known, piecewise continuous scalar function of time and $\beta(\cdot, \cdot)$ is a scalar Brownian motion of diffusion $q(t)$ for all $t \in T = [0, \infty)$, then we want to give meaning to

$$I(t, \cdot) \triangleq \int_{t_0}^t a(\tau)\,d\beta(\tau, \cdot) \tag{4-71}$$

called a *scalar stochastic integral* [1, 3, 5, 13, 14]. The notation provides the insight that for a particular time t, $I(t, \cdot)$ will be a random variable, so that considered as a function of both t and ω, $I(\cdot, \cdot)$ will be a stochastic process. In order to give meaning to this expression, we will require that the Riemann integral $\int_{t_0}^t a(\tau)^2 q(\tau)\,d\tau$ be finite; the need for this assumption will be explained subsequently. Note that we could extend this to the case of stochastic, rather than deterministic, $a(\cdot, \cdot)$ if we desired to develop stochastic integrals appropriate for solutions to *nonlinear* stochastic differential equations; this will be postponed until Chapter 11 (Volume 2).

First partition the time interval $[t_0, t]$ into N steps, not necessarily of equal length, with $t_0 < t_1 < t_2 < \cdots < t_N = t$, and let the maximum time increment be denoted as Δt_N:

$$\Delta t_N = \max_{i = 1, \ldots, N} \{(t_i - t_{i-1})\} \tag{4-72}$$

Now define a special function $a_N(\cdot)$, called a "simple function," through the relation

$$a_N(t) = \begin{cases} a(t_0) & t \in [t_0, t_1) \\ a(t_1) & t \in [t_1, t_2) \\ \vdots & \vdots \\ a(t_N) & t \in [t_{N-1}, t_N) \end{cases} \tag{4-73}$$

This is a piecewise constant approximation to the known function $a(\cdot)$, as depicted in Fig. 4.10. For this simple function, the stochastic integral can be defined constructively as the sum of N increment random variables:

$$I_N(t, \cdot) \triangleq \sum_{i=0}^{N-1} a_N(t_i)[\beta(t_{i+1}, \cdot) - \beta(t_i, \cdot)] \triangleq \int_{t_0}^{t} a_N(\tau) d\beta(\tau, \cdot) \tag{4-74}$$

Let us characterize the random variable $I_N(t, \cdot)$ probabilistically. Since $I_N(t, \cdot)$ is composed of the sum of independent Gaussian increments, $I_N(t, \cdot)$ itself is *Gaussian*. Its mean is

$$E\{I_N(t)\} = E\left\{ \sum_{i=0}^{N-1} a_N(t_i)[\beta(t_{i+1}) - \beta(t_i)] \right\}$$

$$= \sum_{i=0}^{N-1} a_N(t_i)E\{[\beta(t_{i+1}) - \beta(t_i)]\} = 0 \tag{4-75}$$
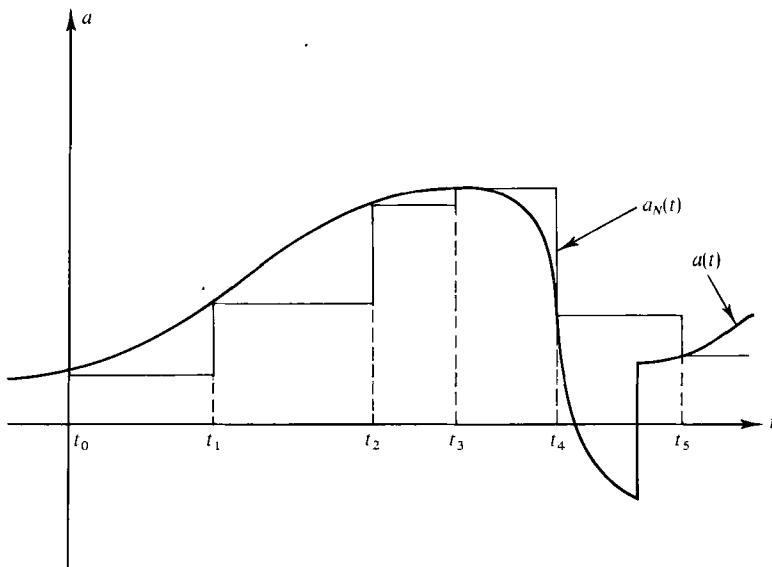
FIG. 4.10   Simple function $a_N(\cdot)$.

which results from $E\{\cdot\}$ being linear and $a_N(\cdot)$ being deterministic (so it can be brought out of the expectations), and then the N separate expectations are zero by the properties of Brownian motion. Its variance can be generated as

$$E\{I_N(t)^2\} = E\left\{\left[\sum_{i=0}^{N-1} a_N(t_i)[\beta(t_{i+1}) - \beta(t_i)]\right]^2\right\}$$

$$= \sum_{i=0}^{N-1} a_N(t_i)^2 E\{[\beta(t_{i+1}) - \beta(t_i)]^2\}$$

where the reduction from N^2 to N separate expectations is due to the independence of Brownian motion increments. Thus,

$$E\{I_N(t)^2\} = \sum_{i=0}^{N-1} a_N(t_i)^2 \int_{t_i}^{t_{i+1}} q(\tau)\,d\tau = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} a_N(t_i)^2 q(\tau)\,d\tau$$

$$= \int_{t_0}^{t} a_N(\tau)^2 q(\tau)\,d\tau \qquad (4\text{-}76)$$

Now it is desired to extend the definition of a stochastic integral in (4-74), valid only for piecewise constant $a_N(\cdot)$, to the case of piecewise continuous $a(\cdot)$. We will consider partitioning the time interval into finer and finer steps, and see if the sequence of random variables $I_N(t, \cdot)$ so formed will converge to some limit as $N \to \infty$.

To motivate this development, consider the set of deterministic functions of time $a(\cdot)$ defined on $[t_0, t]$ such that $\int_{t_0}^{t} a^2(\tau)q(\tau)\,d\tau$ is finite for piecewise continuous $q(\cdot)$, identical to the assumption made at the beginning of this section. On this set of functions, called a *Hilbert space* (or a "complete inner product space"), the "distance" between two functions $a_N(\cdot)$ and $a_P(\cdot)$ can be defined properly by the scalar quantity $\|a_N - a_P\|$ where

$$\|a_N - a_P\|^2 = \int_{t_0}^{t} [a_N(\tau) - a_P(\tau)]^2 q(\tau)\,d\tau \qquad (4\text{-}77)$$

It can be shown that if $a(\cdot)$ is an element of this set, i.e., if $\int_{t_0}^{t} a^2(\tau)q(\tau)\,d\tau$ is finite, then there exists a sequence of simple functions $a_k(\cdot)$ in this set that converges to $a(\cdot)$ as $k \to \infty$ [i.e., as the number of partitions of the time interval goes to ∞ and Δt_N, the maximum time increment, defined in (4-72), goes to zero], where convergence is in the sense that the "distance" between a_k and a converges to zero:

$$\lim_{k \to \infty} \|a - a_k\|^2 = 0 \qquad (4\text{-}78)$$

Moreover, if a sequence of elements from this "Hilbert space" is a "Cauchy" sequence (for an arbitrarily small $\varepsilon > 0$, there exists an integer K such that for all $i > K$ and $j > K$, $\|a_i - a_j\| < \varepsilon$, which heuristically means that the members of the sequence become progressively closer together), then the sequence does

in fact converge to a limit $a(\cdot)$. The limit is itself a member of that Hilbert space, which is assured by the "completeness" of the space.

To consider the convergence of a sequence of stochastic integrals of the form (4-74), define another stochastic integral of this form, but based upon P time partitions, with $P > N$:

$$I_P(t, \cdot) \triangleq \sum_{j=0}^{P-1} a_P(t_j)[\beta(t_{j+1}, \cdot) - \beta(t_j, \cdot)] \tag{4-79}$$

The difference between $I_N(t, \cdot)$ and $I_P(t, \cdot)$ is then

$$[I_N(t) - I_P(t)] = \sum_{i=0}^{N-1} a_N(t_i)[\beta(t_{i+1}) - \beta(t_i)] - \sum_{j=0}^{P-1} a_P(t_j)[\beta(t_{j+1}) - \beta(t_j)]$$

Since $a_N(\cdot)$ and $a_P(\cdot)$ are piecewise constant, their difference must be piecewise constant, with at most $N + P$ points of discontinuity. Thus, for some $K \leq (N + P)$

$$[I_N(t) - I_P(t)] = \sum_{k=0}^{K-1} [a_N(t_k) - a_P(t_k)][\beta(t_{k+1}) - \beta(t_k)]$$

$$\triangleq \int_{t_0}^{t} [a_N(\tau) - a_P(\tau)] \, d\beta(\tau) \tag{4-80}$$

serves to define the integral $\int_{t_0}^{t} [a_N(\tau) - a_P(\tau)] \, d\beta(\tau)$. The mean square value of this difference is

$$E\{[I_N(t) - I_P(t)]^2\} = E\left\{\left[\int_{t_0}^{t} [a_N(\tau) - a_P(\tau)] \, d\beta(\tau)\right]^2\right\}$$

and, since $[a_N(\cdot) - a_P(\cdot)]$ is piecewise constant, this can be shown equal to the *ordinary Riemann integral*:

$$E\{[I_N(t) - I_P(t)]^2\} = \int_{t_0}^{t} [a_N(\tau) - a_P(\tau)]^2 q(\tau) \, d\tau \tag{4-81}$$

Under the assumptions made previously, essentially that the random variables under consideration are zero mean and of finite second moments, we can now speak of the *Hilbert space of random variables* $I(t, \cdot)$, with "distance" between random variables $I_N(t, \cdot)$ and $I_P(t, \cdot)$ defined as $\|I_N(t) - I_P(t)\|$, where

$$\|I_N(t) - I_P(t)\|^2 = E\{[I_N(t) - I_P(t)]^2\} \tag{4-82}$$

Combined with (4-81), this yields a distance measure identical to (4-77). A sequence of random variables $I_1(t, \cdot)$, $I_2(t, \cdot)$, ..., generated by taking finer partitions of $[t_0, t]$ such that Δt_k converges to zero, will be a Cauchy sequence, and thus will converge to a limit in that space, denoted as $I(t, \cdot)$. By Eqs. (4-78)

and (4-82), this assured convergence is in the *mean square sense*:

$$\lim_{k \to \infty} \|\mathsf{I}(t) - \mathsf{I}_k(t)\|^2 = \lim_{k \to \infty} E\{[I(t) - I_k(t)]^2\} = 0 \qquad (4\text{-}83)$$

Thus, we can define the *scalar stochastic integral* $\mathsf{I}(\cdot, \cdot)$ properly through

$$\mathsf{I}(t, \cdot) = \int_{t_0}^{t} a(\tau) \, \mathrm{d}\beta(\tau, \cdot)$$

$$\dot{=} \text{l.i.m.}_{N \to \infty} \mathsf{I}_N(t, \cdot) \triangleq \text{l.i.m.}_{N \to \infty} \int_{t_0}^{t} a_N(\tau) \, \mathrm{d}\beta(\tau) \qquad (4\text{-}84)$$

Since Brownian motion is Gaussian and only linear operations on $\beta(\cdot, \cdot)$ were used in this development, $\mathsf{I}(t, \cdot)$ can be shown to be *Gaussian* with mean and variance

$$E\{\mathsf{I}(t)\} = \lim_{N \to \infty} E\{\mathsf{I}_N(t)\} = 0 \qquad (4\text{-}85\mathrm{a})$$

$$E\{\mathsf{I}(t)^2\} = \lim_{N \to \infty} E\{\mathsf{I}_N(t)^2\} = \int_{t_0}^{t} a(\tau)^2 q(\tau) \, d\tau \qquad (4\text{-}85\mathrm{b})$$

Stochastic integrals exhibit the usual *linear properties* of ordinary integrals:

$$\int_{t_0}^{t_2} a(\tau) \, \mathrm{d}\beta(\tau) = \int_{t_0}^{t_1} a(\tau) \, \mathrm{d}\beta(\tau) + \int_{t_1}^{t_2} a(\tau) \, \mathrm{d}\beta(\tau) \qquad (4\text{-}86\mathrm{a})$$

$$\int_{t_0}^{t} [a(\tau) + a'(\tau)] \, \mathrm{d}\beta(\tau) = \int_{t_0}^{t} a(\tau) \, \mathrm{d}\beta(\tau) + \int_{t_0}^{t} a'(\tau) \, \mathrm{d}\beta(\tau) \qquad (4\text{-}86\mathrm{b})$$

$$\int_{t_0}^{t} a(\tau) \, \mathrm{d} \, [\beta(\tau) + \beta'(\tau)] = \int_{t_0}^{t} a(\tau) \, \mathrm{d}\beta(\tau) + \int_{t_0}^{t} a(\tau) \, \mathrm{d}\beta'(\tau) \qquad (4\text{-}86\mathrm{c})$$

$$\int_{t_0}^{t} c a(\tau) \, \mathrm{d}\beta(\tau) = c \int_{t_0}^{t} a(\tau) \, \mathrm{d}\beta(\tau) = \int_{t_0}^{t} a(\tau) \, \mathrm{d}[c\beta(\tau)] \qquad (4\text{-}86\mathrm{d})$$

Integration by parts is also valid:

$$\int_{t_0}^{t} a(\tau) \, \mathrm{d}\beta(\tau) = a(\tau)\beta(\tau) \Big|_{t_0}^{t} - \int_{t_0}^{t} \beta(\tau) \, da(\tau) \qquad (4\text{-}87)$$

where the last integral term is not a stochastic integral, but an *ordinary* Stieltjes integral definable for each sample of $\beta(\cdot, \cdot)$ if $a(\cdot)$ is of bounded variation.

Viewed as a stochastic process, the stochastic integral can be shown to be *mean square continuous*: $[\mathsf{I}(t_2) - \mathsf{I}(t_1)]$ is zero-mean and

$$E\{[\mathsf{I}(t_2) - \mathsf{I}(t_1)]^2\} = E\left\{\left[\int_{t_0}^{t_2} a(\tau) \, \mathrm{d}\beta(\tau) - \int_{t_0}^{t_1} a(\tau) \, \mathrm{d}\beta(\tau)\right]^2\right\}$$

$$= E\left\{\left[\int_{t_1}^{t_2} a(\tau) \, \mathrm{d}\beta(\tau)\right]^2\right\} = \int_{t_1}^{t_2} a(\tau)^2 q(\tau) \, d\tau \qquad (4\text{-}88)$$

The limit of this ordinary integral as $t_1 \to t_2$ is zero, thereby demonstrating mean square continuity.

Now consider two disjoint intervals $(t_1, t_2]$ and $(t_3, t_4]$, and form

$$[I(t_2) - I(t_1)] = \int_{t_1}^{t_2} a(\tau)\,d\beta(\tau), \qquad [I(t_4) - I(t_3)] = \int_{t_3}^{t_4} a(\tau)\,d\beta(\tau)$$

Since the intervals are disjoint, the independent increments of $\beta(\cdot, \cdot)$ in $(t_1, t_2]$ are independent of the increments in $(t_3, t_4]$. Thus, $[I(t_2) - I(t_1)]$ and $[I(t_4) - I(t_3)]$ are themselves independent, zero-mean, Gaussian increments of the $I(\cdot, \cdot)$ process: *the $I(\cdot, \cdot)$ process is itself a Brownian motion process with rescaled diffusion*, as seen by comparing (4-88) with (4-67).

Extension to the vector case is straightforward. Recall that an s-dimensional vector Brownian motion $\beta(\cdot, \cdot)$ is a Gaussian process composed of independent increments, with statistics

$$E\{\beta(t)\} = 0 \tag{4-69a}$$

$$E\{[\beta(t_2) - \beta(t_1)][\beta(t_2) - \beta(t_1)]^T\} = \int_{t_1}^{t_2} Q(\tau)\,d\tau \tag{4-69b}$$

with the s-by-s diffusion matrix $Q(t)$ symmetric and positive semidefinite and $Q(\cdot)$ a matrix of piecewise continuous functions. If $A(\cdot)$ is an n-by-s matrix of piecewise continuous time functions, then a development analogous to the scalar case yields a definition of an n-dimensional *vector-valued* stochastic integral

$$I(t, \cdot) = \int_{t_0}^{t} A(\tau)\,d\beta(\tau) \tag{4-89}$$

by means of a mean square limit:

$$I(t, \cdot) \triangleq \underset{N \to \infty}{\text{l.i.m.}}\, I_N(t, \cdot) \triangleq \underset{N \to \infty}{\text{l.i.m.}} \int_{t_0}^{t} A_N(\tau)\,d\beta(\tau) \tag{4-90}$$

The random vector $I(t, \cdot)$ is *Gaussian*, with statistics

$$E\{I(t)\} = 0 \tag{4-91a}$$

$$E\{I(t)I^T(t)\} = \int_{t_0}^{t} A(\tau)Q(\tau)A^T(\tau)\,d\tau \tag{4-91b}$$

In such a development, the appropriate "distance" measure $\|I_N(t) - I_P(t)\|$ to replace that defined in (4-82) would be

$$\|I_N(t) - I_P(t)\|^2 = \text{tr}\, E\{[I_N(t) - I_P(t)][I_N(t) - I_P(t)]^T\} \tag{4-92}$$

where tr denotes trace.

Viewed as a function of both $t \in T$ and $\omega \in \Omega$, *the stochastic process $I(\cdot, \cdot)$ is itself a Brownian motion process with rescaled diffusion*:

$$E\{[I(t_2) - I(t_1)][I(t_2) - I(t_1)]^T\} = \int_{t_1}^{t_2} A(\tau)Q(\tau)A^T(\tau)\,d\tau \tag{4-93}$$

4.7   STOCHASTIC DIFFERENTIALS

Given a stochastic integral of the form

$$\mathbf{I}(t) = \mathbf{I}(t_0) + \int_{t_0}^{t} \mathbf{A}(\tau)\,\mathbf{d\beta}(\tau) \tag{4-94}$$

the *stochastic differential* of $\mathbf{I}(t)$ can be defined as

$$\mathbf{dI}(t) = \mathbf{A}(\tau)\,\mathbf{d\beta}(\tau) \tag{4-95}$$

Notice that the differential is defined in terms of the stochastic integral form, and not through an alternate definition in terms of a derivative, since Brownian motion is nondifferentiable. The $\mathbf{dI}(t)$ in (4-95) is thus a differential in the sense that if it is integrated over the entire interval from t_0 to a fixed time t, it yields the random variable $[\mathbf{I}(t) - \mathbf{I}(t_0)]$:

$$\int_{t_0}^{t} \mathbf{dI}(t) = \mathbf{I}(t) - \mathbf{I}(t_0) \tag{4-96}$$

Viewed as a function of t, this yields the stochastic process $[\mathbf{I}(\cdot) - \mathbf{I}(t_0)]$. *Heuristically*, it can be interpreted as an infinitesimal difference

$$\mathbf{dI}(t) = \mathbf{I}(t + dt) - \mathbf{I}(t) \tag{4-97}$$

One particular form required in the next section is the differential of the product of a time function and a stochastic integral. Suppose $\mathbf{s}(\cdot,\cdot)$ is a stochastic integral (which can also be regarded as a Brownian motion) defined through

$$\mathbf{s}(t) = \mathbf{s}(t_0) + \int_{t_0}^{t} \mathbf{A}(\tau)\,\mathbf{d\beta}(\tau) \tag{4-98}$$

Further suppose that $\mathbf{D}(\cdot)$ is a known matrix of differentiable functions, and a random process $\mathbf{y}(\cdot,\cdot)$ were defined by

$$\mathbf{y}(t) = \mathbf{D}(t)\mathbf{s}(t) \tag{4-99}$$

If the time interval $[t_0, t]$ were partitioned into N steps, one could write, assuming $t_{i+1} > t_i$,

$$\mathbf{s}(t) = \mathbf{s}(t_0) + \sum_{i=0}^{N-1} [\mathbf{s}(t_{i+1}) - \mathbf{s}(t_i)], \qquad \mathbf{D}(t) = \mathbf{D}(t_0) + \sum_{i=0}^{N-1} [\mathbf{D}(t_{i+1}) - \mathbf{D}(t_i)]$$

Substituting these back into (4-99) and rearranging yields

$$\mathbf{y}(t) = \sum_{i=0}^{N-1} [\mathbf{D}(t_{i+1}) - \mathbf{D}(t_i)]\mathbf{s}(t_{i+1}) + \sum_{i=0}^{N-1} \mathbf{D}(t_i)[\mathbf{s}(t_{i+1}) - \mathbf{s}(t_i)] + \mathbf{D}(t_0)\mathbf{s}(t_0)$$

Since $\mathbf{D}(\cdot)$ is assumed differentiable, the mean value theorem can be used to write $[\mathbf{D}(t_{i+1}) - \mathbf{D}(t_i)]$ as $\dot{\mathbf{D}}(\tau_i)[t_{i+1} - t_i]$ for some $\tau_i \in (t_i, t_{i+1})$. Putting this into the preceding expression, and taking the mean square limit as $N \to \infty$ yields

$$\mathbf{y}(t) = \int_{t_0}^{t} \dot{\mathbf{D}}(\tau)\mathbf{s}(\tau)\,d\tau + \int_{t_0}^{t} \mathbf{D}(\tau)\,\mathbf{ds}(\tau) + \mathbf{D}(t_0)\mathbf{s}(t_0) \tag{4-100}$$

Note that the first term could be interpreted as an ordinary Riemann integral for each sample function $\mathbf{s}(\cdot, \omega_i)$ of $\mathbf{s}(\cdot, \cdot)$ (i.e., it can be defined in the almost sure sense as well as the mean square sense), but the second term is a stochastic integral which is defined properly only in the mean square sense. From (4-100) and the definition of a stochastic differential, it can be seen that for $\mathbf{y}(t, \cdot)$ given by (4-99),

$$\mathbf{dy}(t) = \dot{\mathbf{D}}(t)\mathbf{s}(t)\,dt + \mathbf{D}(t)\mathbf{ds}(t) \tag{4-101}$$

Equation (4-101) reveals that the stochastic differential of the linear form in (4-99) obeys the same formal rules as the deterministic total differential of a corresponding $\mathbf{y}(t) = \mathbf{D}(t)\mathbf{s}(t)$. This will *not* be the case for nonlinear forms defined in terms of Itô stochastic integrals, as will be seen in Chapter 11 (Volume 2).

4.8   LINEAR STOCHASTIC DIFFERENTIAL EQUATIONS

Section 2.3 developed the solution to linear deterministic state differential equations of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{4-102}$$

with $\mathbf{u}(\cdot)$ a known function of time. Now we would *like* to generate a system model of the form

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{w}(t) \tag{4-103}$$

where $\mathbf{w}(\cdot, \cdot)$ is a white Gaussian noise process of mean zero and strength $\mathbf{Q}(t)$ for all $t \in T$, and analogously develop its solution. (Deterministic driving terms will be admitted subsequently.) However, (4-103) cannot be used rigorously, since its solution cannot be generated.

It *is* possible to write the *linear stochastic differential equation*

$$\mathbf{dx}(t) = \mathbf{F}(t)\mathbf{x}(t)\,dt + \mathbf{G}(t)\,\mathbf{d\beta}(t) \tag{4-104}$$

where $\mathbf{G}(\cdot)$ is a known n-by-s matrix of piecewise continuous functions, and $\boldsymbol{\beta}(\cdot, \cdot)$ is an s-vector-valued Brownian motion process of diffusion $\mathbf{Q}(t)$ for all $t \in T$ [1, 5, 13, 14]. For engineering applications, Eq. (4-103) will often be used to describe a system model, but it is to be interpreted in a rigorous sense as a representation of the more proper relation, (4-104).

Now we seek the solution to (4-104). Recalling the interpretation of stochastic differentials from the last section, we equivalently want to find the $\mathbf{x}(\cdot, \cdot)$ process that satisfies the integral equation

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{F}(\tau)\mathbf{x}(\tau)\,d\tau + \int_{t_0}^{t} \mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \tag{4-105}$$

The last term is a stochastic integral to be understood in the mean square sense; the second term can also be interpreted as a mean square Riemann integral, or

$\int_{t_0}^{t} \mathbf{F}(\tau)\mathbf{x}(\tau, \omega_i)\, d\tau$ can be considered an *ordinary* Riemann integral for a particular sample from the $\mathbf{x}(\cdot, \cdot)$ process.

Let us propose as a solution to (4-104) the process $\mathbf{x}(\cdot, \cdot)$ defined by

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{y}(t) \qquad (4\text{-}106a)$$

where $\mathbf{\Phi}(t, t_0)$ is the state transition matrix that satisfies $\dot{\mathbf{\Phi}}(t, t_0) = \mathbf{F}(t)\mathbf{\Phi}(t, t_0)$ and $\mathbf{\Phi}(t_0, t_0) = \mathbf{I}$, and $\mathbf{y}(t)$ is defined by

$$\mathbf{y}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{\Phi}(t_0, \tau)\mathbf{G}(\tau)\, d\boldsymbol{\beta}(\tau) \qquad (4\text{-}106b)$$

where the order of the time indices in evaluating the preceding $\mathbf{\Phi}(\cdot, \cdot)$ are to be noted (indicative of backward transitions). It must now be demonstrated that this proposed solution satisfies both the initial condition and differential equation (or integral equation). First, it satisfies the initial condition

$$\mathbf{x}(t_0) = \mathbf{\Phi}(t_0, t_0)\mathbf{y}(t_0) = \mathbf{I}\mathbf{x}(t_0) = \mathbf{x}(t_0) \qquad (4\text{-}107)$$

The assumed solution (4-106a) is of the same form as (4-99), so the corresponding $d\mathbf{x}(t)$ can be written from (4-101) as

$$d\mathbf{x}(t) = \frac{d\mathbf{\Phi}(t, t_0)}{dt}\mathbf{y}(t)\, dt + \mathbf{\Phi}(t, t_0)\, d\mathbf{y}(t) \qquad (4\text{-}108)$$

But, from (4-106b) and the definition of a stochastic differential, $d\mathbf{y}(t)$ is just $\mathbf{\Phi}(t_0, t)\mathbf{G}(t)\, d\boldsymbol{\beta}(t)$, so that (4-108) can be used to write

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} d\mathbf{x}(\tau)$$

$$= \mathbf{x}(t_0) + \int_{t_0}^{t} \left[\mathbf{F}(\tau)\mathbf{\Phi}(\tau, t_0)\right]\mathbf{y}(\tau)\, d\tau + \int_{t_0}^{t} \mathbf{\Phi}(\tau, t_0)\mathbf{\Phi}(t_0, \tau)\mathbf{G}(\tau)\, d\boldsymbol{\beta}(\tau)$$

$$= \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{F}(\tau)\mathbf{x}(\tau)\, d\tau + \int_{t_0}^{t} \mathbf{G}(\tau)\, d\boldsymbol{\beta}(\tau) \qquad (4\text{-}109)$$

Thus, the proposed solution form does satisfy the given differential equation and initial condition.

From (4-106), the *solution of the linear stochastic differential equation* (4-104) is given by the stochastic process $\mathbf{x}(\cdot, \cdot)$ defined by

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\, d\boldsymbol{\beta}(\tau) \qquad (4\text{-}110)$$

The extensive development in Sections 4.4–4.7 was required in order to give meaning to this solution form. Note that the Gaussian property of Brownian motion $\boldsymbol{\beta}(\cdot, \cdot)$ was never needed in the development: (4-110) is a valid solution form for any input process having independent increments.

Having obtained the solution as a stochastic process, it is desirable to characterize the statistical properties of that process. First, the *mean* $\mathbf{m}_x(\cdot)$

is described for all $t \in T$ as

$$\mathbf{m}_x(t) = E\{\mathbf{x}(t)\} = \mathbf{\Phi}(t, t_0)E\{\mathbf{x}(t_0)\} + E\left\{\int_{t_0}^t \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau)\right\}$$

$$\mathbf{m}_x(t) = \mathbf{\Phi}(t, t_0)\mathbf{m}_x(t_0) \qquad \qquad \text{(4-111)}$$

since the stochastic integral is of mean zero. To obtain the mean squared value of $\mathbf{x}(t)$, Eq. (4-110) can be used to write $E\{\mathbf{x}(t)\mathbf{x}^T(t)\}$ as the sum of four separate expectations. However, Brownian motion is implicitly independent of $\mathbf{x}(t_0)$ by its definition, so the two cross terms are $\mathbf{0}$, such as:

$$E\left\{\left[\int_{t_0}^t \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau)\right][\mathbf{x}^T(t_0)]\right\} = E\left\{\int_{t_0}^t \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau)\right\}E\{\mathbf{x}^T(t_0)\} = \mathbf{0}$$

Thus, the *mean squared value* or *correlation matrix* of $\mathbf{x}(t)$ is

$$E\{\mathbf{x}(t)\mathbf{x}^T(t)\} = \mathbf{\Phi}(t, t_0)E\{\mathbf{x}(t_0)\mathbf{x}^T(t_0)\}\mathbf{\Phi}^T(t, t_0)$$

$$+ \int_{t_0}^t \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\mathbf{\Phi}^T(t, \tau)\,d\tau \qquad \text{(4-112)}$$

where $\mathbf{Q}(t)$ is the diffusion of the Brownian motion $\mathbf{\beta}(\cdot, \cdot)$ at time t. The form of the *ordinary* Riemann integral in this expression is derived directly from Eq. (4-93).

The *covariance* can be derived directly from the mean square value by substituting

$$E\{\mathbf{x}(t)\mathbf{x}^T(t)\} = \mathbf{P}_{xx}(t) + \mathbf{m}_x(t)\mathbf{m}_x^T(t) \qquad \text{(4-113a)}$$

$$E\{\mathbf{x}(t_0)\mathbf{x}^T(t_0)\} = \mathbf{P}_{xx}(t_0) + \mathbf{m}_x(t_0)\mathbf{m}_x^T(t_0) \qquad \text{(4-113b)}$$

into (4-112) and incorporating (4-111) to yield

$$\mathbf{P}_{xx}(t) = \mathbf{\Phi}(t, t_0)\mathbf{P}_{xx}(t_0)\mathbf{\Phi}^T(t, t_0) + \int_{t_0}^t \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\mathbf{\Phi}^T(t, \tau)\,d\tau \quad \text{(4-114)}$$

From (4-110) it can be seen that, if $\mathbf{x}(t_0)$ is a Gaussian random variable or if it is nonrandom (known exactly), then $\mathbf{x}(t)$ for any fixed t is a *Gaussian* random variable. Thus, the first order density $f_{\mathbf{x}(t)}(\mathbf{\xi})$ is completely determined by the mean and covariance in (4-111) and (4-114) as

$$f_{\mathbf{x}(t)}(\mathbf{\xi}) = \left[(2\pi)^{n/2}|\mathbf{P}_{xx}(t)|^{1/2}\right]^{-1}\exp\left\{-\tfrac{1}{2}[\mathbf{\xi} - \mathbf{m}_x(t)]^T\mathbf{P}_{xx}^{-1}(t)[\mathbf{\xi} - \mathbf{m}_x(t)]\right\} \quad \text{(4-115)}$$

Moreover, because the stochastic integral in (4-110) is composed of independent Gaussian increments, $\mathbf{x}(\cdot, \cdot)$ is a *Gaussian process*. Thus, the joint density $f_{\mathbf{x}(t_1), \mathbf{x}(t_2), \dots, \mathbf{x}(t_N)}(\mathbf{\xi}_1, \mathbf{\xi}_2, \dots, \mathbf{\xi}_N)$ is a Gaussian density for any choice of $t_1, t_2, \dots, t_N$. Its mean components and covariance block-diagonal terms are depicted by (4-111) and (4-114) for $t = t_1, t_2, \dots, t_N$. To completely specify this density requires an expression for the covariance kernel $\mathbf{P}_{xx}(t_i, t_j)$, to be derived next.

For $t_2 \geq t_1 \geq t_0$, $\mathbf{x}(t_2)$ can be written as

$$
\begin{aligned}
\mathbf{x}(t_2) &= \boldsymbol{\Phi}(t_2, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t_2} \boldsymbol{\Phi}(t_2, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \\
&= \boldsymbol{\Phi}(t_2, t_1)\boldsymbol{\Phi}(t_1, t_0)\mathbf{x}(t_0) + \boldsymbol{\Phi}(t_2, t_1)\int_{t_0}^{t_1} \boldsymbol{\Phi}(t_1, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \\
&\quad + \int_{t_1}^{t_2} \boldsymbol{\Phi}(t_2, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \\
&= \boldsymbol{\Phi}(t_2, t_1)\mathbf{x}(t_1) + \int_{t_1}^{t_2} \boldsymbol{\Phi}(t_2, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \qquad (4\text{-}116)
\end{aligned}
$$

Since the increments of $\boldsymbol{\beta}(\cdot, \cdot)$ over $[t_1, t_2)$ are independent of both the increments over $[t_0, t_1)$ and $\mathbf{x}(t_0)$, the *autocorrelation* $E\{\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_1)\}$ can be written as

$$
\begin{aligned}
E\{\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_1)\} &= \boldsymbol{\Phi}(t_2, t_1)E\{\mathbf{x}(t_1)\mathbf{x}^{\mathrm{T}}(t_1)\} + E\left\{\int_{t_1}^{t_2} \boldsymbol{\Phi}(t_2, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau)\mathbf{x}^{\mathrm{T}}(t_1)\right\} \\
&= \boldsymbol{\Phi}(t_2, t_1)E\{\mathbf{x}(t_1)\mathbf{x}^{\mathrm{T}}(t_1)\} \qquad (4\text{-}117)
\end{aligned}
$$

Then (4-113) can be used to show that the desired *covariance kernel* for $t_2 \geq t_1$ is

$$
\mathbf{P}_{xx}(t_2, t_1) = \boldsymbol{\Phi}(t_2, t_1)\mathbf{P}_{xx}(t_1, t_1) = \boldsymbol{\Phi}(t_2, t_1)\mathbf{P}_{xx}(t_1) \qquad (4\text{-}118)
$$

Close inspection of Eq. (4-116) reveals the fact that $\mathbf{x}(\cdot, \cdot)$ is not only a Gaussian process, but a *Gauss–Markov process* as described in Section 4.4. The probability law that describes the process evolution in the future depends only on the present process description (at time t_1 for instance) and not upon the history of the process evolution (before time t_1).

Equations (4-111), (4-114), and (4-118) are the fundamental characterization of the Gauss–Markov process solution (4-110) to the linear stochastic differential equation (4-104). However, it is often convenient to utilize the equivalent set of *differential equations for* $\mathbf{m}_x(t)$ *and* $\mathbf{P}_{xx}(t)$ to describe their evolution in time. Differentiating (4-111) yields the *mean time propagation* as

$$
\begin{aligned}
\dot{\mathbf{m}}_x(t) &= \dot{\boldsymbol{\Phi}}(t, t_0)\mathbf{m}_x(t_0) = \mathbf{F}(t)\boldsymbol{\Phi}(t, t_0)\mathbf{m}_x(t_0) \\
\dot{\mathbf{m}}_x(t) &= \mathbf{F}(t)\mathbf{m}_x(t) \qquad (4\text{-}119)
\end{aligned}
$$

Since the stochastic driving term in (4-104) has zero-mean, the mean of $\mathbf{x}(t)$ satisfies the homogeneous form of the state equation. Differentiating (4-114) yields, using Leibnitz' rule,

$$
\begin{aligned}
\dot{\mathbf{P}}_{xx}(t) &= \mathbf{F}(t)\boldsymbol{\Phi}(t, t_0)\mathbf{P}_{xx}(t_0)\boldsymbol{\Phi}^{\mathrm{T}}(t, t_0) + \boldsymbol{\Phi}(t, t_0)\mathbf{P}_{xx}(t_0)\boldsymbol{\Phi}^{\mathrm{T}}(t, t_0)\mathbf{F}^{\mathrm{T}}(t) \\
&\quad + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) + \int_{t_0}^{t} \mathbf{F}(t)\boldsymbol{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\boldsymbol{\Phi}^{\mathrm{T}}(t, \tau)\,d\tau \\
&\quad + \int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\boldsymbol{\Phi}^{\mathrm{T}}(t, \tau)\mathbf{F}^{\mathrm{T}}(t)\,d\tau
\end{aligned}
$$

Taking $\mathbf{F}(t)$ and $\mathbf{F}^T(t)$ out of the integrals since they are not functions of τ, and rearranging, yields

$$\dot{\mathbf{P}}_{xx}(t) = \mathbf{F}(t)\mathbf{P}_{xx}(t) + \mathbf{P}_{xx}(t)\mathbf{F}^T(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^T(t) \qquad (4\text{-}120)$$

These are general relationships in that they allow time-varying system models and Brownian motion diffusion as well as time-invariant parameters.
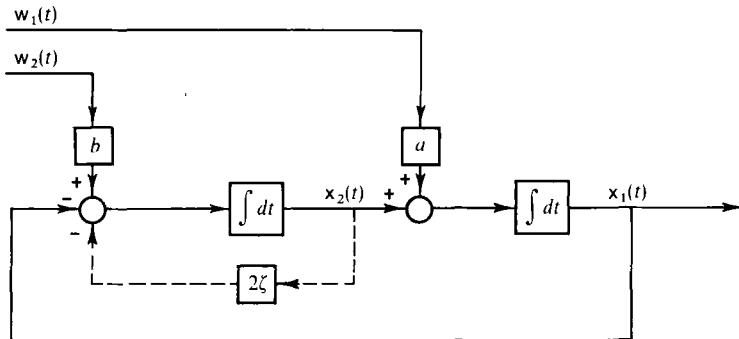
FIG. 4.11   Second order system model.

EXAMPLE 4.5   Consider a 1 rad/sec oscillator driven by white Gaussian noises, as depicted in Fig. 4.11 for $\zeta = 0$. The state equations can be written in the nonrigorous white noise notation as

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}\begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}$$

$$\dot{\mathbf{x}}(t) \quad = \quad \mathbf{F} \quad \mathbf{x}(t) \quad + \quad \mathbf{G} \quad \mathbf{w}(t)$$

Suppose the initial conditions at time $t = 0$ are that the oscillator is known to start precisely at $x_1(0) = 1$, $x_2(0) = 3$. Let $w_1(\cdot)$ and $w_2(\cdot)$ be independent, zero-mean, and of strength one and two, respectively:

$$E\{w_1(t)w_1(t + \tau)\} = 1\delta(\tau), \qquad E\{w_2(t)w_2(t + \tau)\} = 2\delta(\tau), \qquad E\{w_1(t)w_2(t + \tau)\} = 0$$

Now we want to derive expressions for $\mathbf{m}_x(t)$ and $\mathbf{P}_{xx}(t)$ for all $t \geq 0$.

Since the initial conditions are known without uncertainty, $\mathbf{x}(t_0)$ can be modeled as a Gaussian random variable with zero covariance:

$$\mathbf{m}_x(0) = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \qquad \mathbf{P}_{xx}(0) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Furthermore, from the given information, $\mathbf{Q}$ can be identified as

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

To use (4-111) requires knowledge of the state transition matrix, the solution to $\dot{\mathbf{\Phi}}(t, t_0) = \mathbf{F}(t)\mathbf{\Phi}(t, t_0)$, $\mathbf{\Phi}(t_0, t_0) = \mathbf{I}$. Since the system is time invariant, Laplace transform techniques could be used also. The result is

$$\mathbf{\Phi}(t, t_0) = \mathbf{\Phi}(t - t_0) = \begin{bmatrix} \cos(t - t_0) & \sin(t - t_0) \\ -\sin(t - t_0) & \cos(t - t_0) \end{bmatrix}$$

Thus, (4-111) yields $\mathbf{m}_x(t)$ as

$$\mathbf{m}_x(t) = \mathbf{\Phi}(t,0)\mathbf{m}_x(0) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}\begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} \cos t + 3\sin t \\ -\sin t + 3\cos t \end{bmatrix}$$

$\mathbf{P}_{xx}(t)$ is generated from (4-114) as

$$\mathbf{P}_{xx}(t) = \int_0^t \mathbf{\Phi}(t,\tau)\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}(t,\tau)\,d\tau$$

$$= \int_0^t \begin{bmatrix} \cos(t-\tau) & \sin(t-\tau) \\ -\sin(t-\tau) & \cos(t-\tau) \end{bmatrix}\begin{bmatrix} a^2 & 0 \\ 0 & 2b^2 \end{bmatrix}\begin{bmatrix} \cos(t-\tau) & -\sin(t-\tau) \\ \sin(t-\tau) & \cos(t-\tau) \end{bmatrix}d\tau$$

$$= \begin{bmatrix} \left(\dfrac{a^2+2b^2}{2}\right)t + \left(\dfrac{a^2-2b^2}{4}\right)\sin 2t & \left(\dfrac{2b^2-a^2}{2}\right)\sin^2 t \\[3mm] \left(\dfrac{2b^2-a^2}{2}\right)\sin^2 t & \left(\dfrac{a^2+2b^2}{2}\right)t + \left(\dfrac{2b^2-a^2}{4}\right)\sin 2t \end{bmatrix}$$

Note that the covariance is diverging: the diagonal terms grow linearly with time with a sinusoid superimposed. Thus a single sample from the $\mathbf{x}(\cdot,\cdot)$ process would be expected to be divergent as well as oscillatory. ∎

EXAMPLE 4.6   Consider the same second order system as in Fig. 4.11, but with damping added by letting the damping ratio ζ be nonzero. $\mathbf{F}$ then becomes

$$\mathbf{F} = \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix}$$

and the same calculations can be performed, or (4-119) and (4-120) used to write

$$\dot{\mathbf{m}}_x(t) = \begin{bmatrix} \dot{m}_{x_1}(t) \\ \dot{m}_{x_2}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix}\begin{bmatrix} m_{x_1}(t) \\ m_{x_2}(t) \end{bmatrix} = \begin{bmatrix} m_{x_2}(t) \\ -m_{x_1}(t) - 2\zeta m_{x_2}(t) \end{bmatrix}$$

$$\dot{\mathbf{P}}_{xx}(t) = \begin{bmatrix} \dot{P}_{11}(t) & \dot{P}_{12}(t) \\ \dot{P}_{12}(t) & \dot{P}_{22}(t) \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix}\begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{12}(t) & P_{22}(t) \end{bmatrix} + \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{12}(t) & P_{22}(t) \end{bmatrix}\begin{bmatrix} 0 & -1 \\ 1 & -2\zeta \end{bmatrix} + \begin{bmatrix} a^2 & 0 \\ 0 & 2b^2 \end{bmatrix}$$

$$= \begin{bmatrix} 2P_{12}(t) + a^2 & -P_{11}(t) - 2\zeta P_{12}(t) + P_{22}(t) \\ -P_{11}(t) - 2\zeta P_{12}(t) + P_{22}(t) & -2P_{12}(t) - 4\zeta P_{22}(t) + 2b^2 \end{bmatrix}$$

This covariance does *not* grow without bound, and a steady state value can be found by evaluating $\dot{\mathbf{P}}_{xx}(t) = \mathbf{0}$, to yield

$$\mathbf{P}_{xx}(t \to \infty) = \begin{bmatrix} \dfrac{a^2(1-4\zeta^2)+2b^2}{4\zeta} & -\dfrac{a^2}{2} \\[4mm] -\dfrac{a^2}{2} & \dfrac{a^2+2b^2}{4\zeta} \end{bmatrix}$$ ∎

Deterministic control inputs can be added to the system model (4-103) or (4-104) without contributing any substantial complexity to the previous develop-

ment. Let the linear stochastic differential equation be written as

$$\mathbf{dx}(t) = \mathbf{F}(t)\mathbf{x}(t)\,dt + \mathbf{B}(t)\mathbf{u}(t)\,dt + \mathbf{G}(t)\,\mathbf{d\beta}(t) \qquad (4\text{-}121a)$$

or, in the less rigorous white noise notation,

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \qquad (4\text{-}121b)$$

where $\mathbf{u}(t)$ is an r-dimensional vector of deterministic control inputs applied at time t, and $\mathbf{B}(t)$ is an n-by-r control input matrix. The solution to (4-121) is the process $\mathbf{x}(\cdot,\cdot)$ defined by

$$\mathbf{x}(t) = \mathbf{\Phi}(t,t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau + \int_{t_0}^{t} \mathbf{\Phi}(t,\tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \quad (4\text{-}122)$$

The only difference between this and (4-110) is the addition of the ordinary Riemann integral $\int_{t_0}^{t} \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$, a *known n-vector* for fixed time t. Since this contributes no additional uncertainty, *only the mean of $\mathbf{x}(t)$ is affected* by this addition: $\mathbf{P}_{xx}(t)$ is still propagated by (4-114) or (4-120), while $\mathbf{m}_x(t)$ is propagated by

$$\mathbf{m}_x(t) = \mathbf{\Phi}(t,t_0)\mathbf{m}_x(t_0) + \int_{t_0}^{t} \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \qquad (4\text{-}123a)$$

or

$$\dot{\mathbf{m}}_x(t) = \mathbf{F}(t)\mathbf{m}_x(t) + \mathbf{B}(t)\mathbf{u}(t) \qquad (4\text{-}123b)$$

EXAMPLE 4.7 Consider Example 4.6 but with $w_1(t)$ changed to $[u(t) + w_1(t)]$. where $u(t) = u = $ constant for all $t \geq 0$. Then the state equation can be written as

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} a \\ 0 \end{bmatrix} u + \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}$$

so that $\dot{\mathbf{m}}_x(t)$ is given by

$$\begin{bmatrix} \dot{m}_{x_1}(t) \\ \dot{m}_{x_2}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix} \begin{bmatrix} m_{x_1}(t) \\ m_{x_2}(t) \end{bmatrix} + \begin{bmatrix} a \\ 0 \end{bmatrix} u = \begin{bmatrix} m_{x_2}(t) + au \\ -m_{x_1}(t) - 2\zeta m_{x_2}(t) \end{bmatrix}$$

There is a steady state value of $\mathbf{m}_x(t)$. and it can be found by setting $\dot{\mathbf{m}}_x(t) = 0$. Unlike Example 4.6, for which $\mathbf{m}_x(t \to \infty) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. this yields

$$\mathbf{m}_x(t \to \infty) = \begin{bmatrix} 2\zeta au \\ -au \end{bmatrix}$$

The covariance relations remain unchanged from Example 4.6. ■

The results of this section can be obtained using *formal procedures* on linear differential equations driven by white Gaussian noise. If $\mathbf{\beta}(\cdot,\cdot)$ is Brownian motion of diffusion $\mathbf{Q}(t)$ for all time $t \in T$, then a stochastic integral $\mathbf{I}(t,\cdot) = \int_{t_0}^{t} \mathbf{A}(\tau)\mathbf{d\beta}(\tau,\cdot)$ can be defined properly in the mean square sense, with mean zero and

$$E\{\mathbf{I}(t)\mathbf{I}^T(t)\} = \int_{t_0}^{t} \mathbf{A}(\tau)\mathbf{Q}(\tau)\mathbf{A}^T(\tau)\,d\tau$$

as found previously. If the *formal* (nonexistent) derivative of the Brownian motion is taken, white Gaussian noise $\mathbf{w}(\cdot,\cdot)$ results, with mean zero and strength $\mathbf{Q}(t)$: $E\{\mathbf{w}(t)\mathbf{w}^\mathsf{T}(t')\} = \mathbf{Q}(t)\,\delta(t - t')$. The results gained by assuming the existence of this formal derivative will be *consistent* with the results that were derived *properly*. Based on this formal definition, we could write

$$\mathbf{l}(t) \stackrel{\mathrm{f}}{=} \int_{t_0}^{t} \mathbf{A}(\tau)\mathbf{w}(\tau)\,d\tau$$

where $\stackrel{\mathrm{f}}{=}$ denotes "formally." Then the mean and covariance would be

$$E\{\mathbf{l}(t)\} \stackrel{\mathrm{f}}{=} \int_{t_0}^{t} \mathbf{A}(\tau)E\{\mathbf{w}(\tau)\}\,d\tau = \mathbf{0}$$

$$E\{\mathbf{l}(t)\mathbf{l}^\mathsf{T}(t)\} \stackrel{\mathrm{f}}{=} E\left\{\left[\int_{t_0}^{t} \mathbf{A}(\tau_1)\mathbf{w}(\tau_1)\,d\tau_1\right]\left[\int_{t_0}^{t} \mathbf{A}(\tau_2)\mathbf{w}(\tau_2)\,d\tau_2\right]^\mathsf{T}\right\}$$

$$\stackrel{\mathrm{f}}{=} \int_{t_0}^{t}\int_{t_0}^{t} \mathbf{A}(\tau_1)E\{\mathbf{w}(\tau_1)\mathbf{w}^\mathsf{T}(\tau_2)\}\mathbf{A}^\mathsf{T}(\tau_2)\,d\tau_2\,d\tau_1$$

$$\stackrel{\mathrm{f}}{=} \int_{t_0}^{t}\left[\int_{t_0}^{t} \mathbf{A}(\tau_1)\mathbf{Q}(\tau_1)\,\delta(\tau_1 - \tau_2)\mathbf{A}^\mathsf{T}(\tau_2)\,d\tau_2\right]d\tau_1$$

$$\stackrel{\mathrm{f}}{=} \int_{t_0}^{t} \mathbf{A}(\tau_1)\mathbf{Q}(\tau_1)\mathbf{A}^\mathsf{T}(\tau_1)\,d\tau_1$$

so, at least formally, this is consistent. In terms of this white noise notation, the state equation can be written as in (4-121b) and the solution written formally as in (4-122), with identical statistical characteristics.

What is gained by avoiding the simplistic approach? First, such an approach does not force one to ask himself some fundamental questions, the answers to which provide significant insights into the nature of stochastic processes themselves. Second, basing an entire development of estimators and controllers upon an improperly defined model will make the validity of everything that follows subject to doubt. Finally, such an approach is totally misleading, in that when nonlinear stochastic differential equations are considered, formal procedures *will not* provide results consistent with those obtained properly through the Itô stochastic integral.

4.9 LINEAR STOCHASTIC DIFFERENCE EQUATIONS

Consider the concept of an *equivalent discrete-time system model* motivated by eventual digital computer implementations of algorithms, as introduced in Section 2.4. Suppose we obtain discrete-time measurements from a continuous-time system described by Eq. (4-121), with $\mathbf{u}(t)$ held constant over each sample period from sample time t_i to t_{i+1}. At the discrete time t_{i+1}, the solution can be

written as

$$\mathbf{x}(t_{i+1}) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \left[\int_{t_i}^{t_{i+1}} \mathbf{\Phi}(t_{i+1}, \tau)\mathbf{B}(\tau)\, d\tau\right]\mathbf{u}(t_i)$$

$$+ \left[\int_{t_i}^{t_{i+1}} \mathbf{\Phi}(t_{i+1}, \tau)\mathbf{G}(\tau)\, d\boldsymbol{\beta}(\tau)\right] \tag{4-124}$$

This can be written as an equivalent stochastic difference equation, i.e., an equivalent discrete-time model, as:

$$\mathbf{x}(t_{i+1}) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \mathbf{B_d}(t_i)\mathbf{u}(t_i) + \mathbf{w_d}(t_i) \tag{4-125}$$

where $\mathbf{B_d}(t_i)$ is the discrete-time input matrix defined by

$$\mathbf{B_d}(t_i) = \int_{t_i}^{t_{i+1}} \mathbf{\Phi}(t_{i+1}, \tau)\mathbf{B}(\tau)\, d\tau \tag{4-126}$$

and $\mathbf{w_d}(\cdot, \cdot)$ is an n-vector-valued white Gaussian discrete-time stochastic process with statistics duplicating those of $\int_{t_i}^{t_{i+1}} \mathbf{\Phi}(t_{i+1}, \tau)\mathbf{G}(\tau)\, d\boldsymbol{\beta}(\tau)$ for all $t_i \in T$,

$$E\{\mathbf{w_d}(t_i)\} = \mathbf{0} \tag{4-127a}$$

$$E\{\mathbf{w_d}(t_i)\mathbf{w_d}^{\mathrm{T}}(t_i)\} = \mathbf{Q_d}(t_i) = \int_{t_i}^{t_{i+1}} \mathbf{\Phi}(t_{i+1}, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\mathbf{\Phi}^{\mathrm{T}}(t_{i+1}, \tau)\, d\tau \tag{4-127b}$$

$$E\{\mathbf{w_d}(t_i)\mathbf{w_d}^{\mathrm{T}}(t_j)\} = \mathbf{0}, \qquad t_i \neq t_j \tag{4-127c}$$

Thus, (4-125) defines a discrete-time stochastic process which has identical characteristics to the result of sampling the solution process to (4-121) at the discrete times $t_0, t_1, t_2, \ldots$. The subscript d denotes discrete-time, to avoid confusion between $\mathbf{B}(\cdot)$ and $\mathbf{B_d}(\cdot)$, etc.

Computationally, it is often more convenient to specify differential equations to solve over an interval rather than an integral relation as in (4-126) or (4-127b). To accomplish this, first define for any $t \in [t_i, t_{i+1})$

$$\bar{\mathbf{B}}(t, t_i) = \int_{t_i}^{t} \mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\, d\tau \tag{4-128a}$$

$$\bar{\mathbf{Q}}(t, t_i) = \int_{t_i}^{t} \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\mathbf{\Phi}^{\mathrm{T}}(t, \tau)\, d\tau \tag{4-128b}$$

Taking the time derivative of these relations yields the desired result: the differential equations to be solved over each interval $[t_i, t_{i+1}) \in T$ to generate $\mathbf{\Phi}(t_{i+1}, t_i)$, $\mathbf{B_d}(t_i)$, and $\mathbf{Q_d}(t_i)$, which completely describe the equivalent discrete-time model (4-125) corresponding to (4-121) [12]:

$$\dot{\mathbf{\Phi}}(t, t_i) = \mathbf{F}(t)\mathbf{\Phi}(t, t_i) \tag{4-129a}$$

$$\dot{\bar{\mathbf{B}}}(t, t_i) = \mathbf{F}(t)\bar{\mathbf{B}}(t, t_i) + \mathbf{B}(t) \tag{4-129b}$$

$$\dot{\bar{\mathbf{Q}}}(t, t_i) = \mathbf{F}(t)\bar{\mathbf{Q}}(t, t_i) + \bar{\mathbf{Q}}(t, t_i)\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) \tag{4-129c}$$

These are integrated forward from the initial conditions

$$\boldsymbol{\Phi}(t_i, t_i) = \mathbf{I}; \qquad \bar{\mathbf{B}}(t_i, t_i) = \mathbf{0}; \qquad \bar{\mathbf{Q}}(t_i, t_i) = \mathbf{0} \qquad (4\text{-}130)$$

to the time t_{i+1}, to yield the desired $\boldsymbol{\Phi}(t_{i+1}, t_i)$ and

$$\mathbf{B}_d(t_i) \triangleq \bar{\mathbf{B}}(t_{i+1}, t_i) \qquad (4\text{-}131a)$$

$$\mathbf{Q}_d(t_i) \triangleq \bar{\mathbf{Q}}(t_{i+1}, t_i) \qquad (4\text{-}131b)$$

In the general case, these integrations must be carried out separately for each sample period. However, many practical cases involve time-invariant system models and stationary noise inputs, for which a single set of integrations suffices for all sample periods. Moreover, for time-invariant or slowly varying $\mathbf{F}(\cdot)$, $\mathbf{B}(\cdot)$, and $[\mathbf{G}(\cdot)\mathbf{Q}(\cdot)\mathbf{G}^T(\cdot)]$, if the sample period is short compared to the system's natural transients, a first order approximation to the solution of (4-129)–(4-131) can often be used [12]; namely,

$$\boldsymbol{\Phi}(t_{i+1}, t_i) \cong \mathbf{I} + \mathbf{F}(t_i)[t_{i+1} - t_i] \qquad (4\text{-}132a)$$

$$\mathbf{B}_d(t_i) \cong \mathbf{B}(t_i)[t_{i+1} - t_i] \qquad (4\text{-}132b)$$

$$\mathbf{Q}_d(t_i) \cong \mathbf{G}(t_i)\mathbf{Q}(t_i)\mathbf{G}^T(t_i)[t_{i+1} - t_i] \qquad (4\text{-}132c)$$

Equation (4-125) is a particular case of a *linear stochastic difference equation* of the general form

$$\mathbf{x}(t_{i+1}) = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \mathbf{B}_d(t_i)\mathbf{u}(t_i) + \mathbf{G}_d(t_i)\mathbf{w}_d(t_i) \qquad (4\text{-}133)$$

where $\mathbf{w}_d(\cdot, \cdot)$ is an s-vector-valued discrete-time white Gaussian noise process, with mean zero and covariance kernel

$$E\{\mathbf{w}_d(t_i)\mathbf{w}_d{}^T(t_j)\} = \begin{cases} \mathbf{Q}_d(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \qquad (4\text{-}134)$$

and $\mathbf{G}_d(t_i)$ is an n-by-s noise input matrix for all $t_i \in T$. In (4-134), $\mathbf{Q}_d(t_i)$ is a real, symmetric, positive semidefinite s-by-s matrix for all $t_i \in T$. Sometimes difference equations are written in terms of the argument i, for instant, rather than time t_i. Recall from Section 2.4 that if (4-133) did not arise from discretizing a a continuous-time model, there is no longer any assurance that $\boldsymbol{\Phi}(t_{i+1}, t_i)$ is always nonsingular.

The mean and covariance of the $\mathbf{x}(\cdot, \cdot)$ process defined by (4-133) propagate as

$$\mathbf{m}_x(t_{i+1}) = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{m}_x(t_i) + \mathbf{B}_d(t_i)\mathbf{u}(t_i) \qquad (4\text{-}135a)$$

$$\mathbf{P}_{xx}(t_{i+1}) = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{P}_{xx}(t_i)\boldsymbol{\Phi}^T(t_{i+1}, t_i) + \mathbf{G}_d(t_i)\mathbf{Q}_d(t_i)\mathbf{G}_d{}^T(t_i) \qquad (4\text{-}135b)$$

EXAMPLE 4.8   Consider a digital simulation of a first order lag as depicted in Fig. 4.12. Let $w(\cdot, \cdot)$ be a white Gaussian noise with zero mean and

$$E[\mathbf{w}(t)\mathbf{w}(t + \tau)] = Q\,\delta(\tau)$$
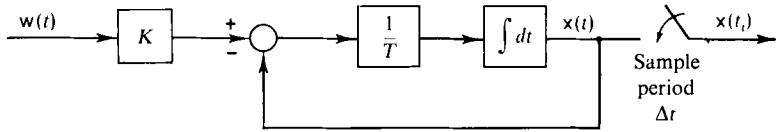
FIG. 4.12   First order lag system model.

and let the interval between sample times be a constant, $(t_{i+1} - t_i) = \Delta t$. From the figure, the state equation can be written as

$$\dot{x}(t) = -(1/T)x(t) + (K/T)w(t)$$

Now we desire an equivalent discrete-time model. The state transition matrix is

$$\Phi(t, \tau) = \Phi(t - \tau) = e^{-(t-\tau)/T}$$

Therefore, the desired model is defined by

$$x(t_{i+1}) = e^{-\Delta t/T}x(t_i) + w_d(t_i)$$

where $w_d(\cdot, \cdot)$ is a white Gaussian discrete-time process with mean zero and

$$E\{w_d(t_i)^2\} = \int_{t_i}^{t_{i+1}} \Phi^2(t_{i+1}, \tau)\left[\frac{K}{T}\right]^2 Q \, d\tau = \frac{QK^2}{2T}\left[1 - e^{-2\Delta t/T}\right] = Q_d$$

Steady state performance is reached, and can be found from either the continuous-time or discrete-time model. For the continuous-time model, in general, set

$$\dot{\mathbf{P}}(t) = \mathbf{F}\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T = 0$$

which here becomes

$$\dot{P}(t) = -(2/T)P(t) + (K^2Q/T^2) = 0$$

so that

$$P(t \to \infty) = QK^2/(2T)$$

For the discrete-time model, the same result can be obtained by setting

$$\mathbf{P}(t_{i+1}) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i)\mathbf{\Phi}^T(t_{i+1}, t_i) + \mathbf{G}_d\mathbf{Q}_d\mathbf{G}_d{}^T = \mathbf{P}(t_i)$$

or

$$P(t_{i+1}) = e^{-2\Delta t/T}P(t_i) + (QK^2/(2T))\left[1 - e^{-2\Delta t/T}\right] = P(t_i) = P$$

so that

$$P\left[1 - e^{-2\Delta t/T}\right] = (QK^2/(2T))\left[1 - e^{-2\Delta t/T}\right]$$

or

$$P = QK^2/(2T)$$

Assume that $m_x(t_0) = 0$, so that

$$E\{x(t_{i+1})x(t_i)\} = \Phi(t_{i+1}, t_i)P(t_i) = e^{-\Delta t/T}P(t_i)$$

which converges to $e^{-\Delta t/T}P$ in steady state. Thus, if the sample period Δt is long compared to the first order lag time constant T, then there is very little correlation between $x(t_{i+1})$ and $x(t_i)$.   ∎

Monte Carlo simulations of systems will be discussed in detail in Section 6.8. Moreover, Problem 7.14 will describe means of generating samples of $\mathbf{w}_d(\cdot,\cdot)$, required for simulations of (4-133) for the general case in which $\mathbf{Q}_d(t_i)$ in (4-134) is nondiagonal.

4.10   THE OVERALL SYSTEM MODEL

The previous two sections developed continuous-time and discrete-time state propagation models. To develop an overall system model, the measurements available from a system must be described. We will be interested mostly in data samples rather than continuously available outputs. At time t_i, the measurements can be described through the m-dimensional random vector $\mathbf{z}(t_i,\cdot)$:

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{4-136}$$

where the number of measurements m is typically smaller than the state dimension n, $\mathbf{H}(t_i)$ is an m-by-n measurement matrix, and $\mathbf{v}(t_i)$ is an m-dimensional vector of additive noise. Thus, each of the m measurements available at time t_i is assumed to be expressible as a linear combination of state variables, corrupted by noise. The physical data (i.e., numbers) from measuring devices are then *realizations* of (4-136):

$$\mathbf{z}(t_i) = \mathbf{z}(t_i,\omega_k) = \mathbf{H}(t_i)\mathbf{x}(t_i,\omega_k) + \mathbf{v}(t_i,\omega_k) \tag{4-137}$$

The noise $\mathbf{v}(\cdot,\cdot)$ will be modeled as a white Gaussian discrete-time stochastic process, with

$$E\{\mathbf{v}(t_i)\} = \mathbf{0} \tag{4-138a}$$

$$E\{\mathbf{v}(t_i)\mathbf{v}^T(t_j)\} = \begin{cases} \mathbf{R}(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{4-138b}$$

It will also be assumed that $\mathbf{v}(t_i,\cdot)$ is independent of both the initial condition $\mathbf{x}(t_0)$ and the dynamic driving noise $\boldsymbol{\beta}(t_j,\cdot)$ or $\mathbf{w}_d(t_j,\cdot)$ for all $t_i, t_j \in T$. A generalization allowing correlation between these various random variables is possible, but this will be pursued later.

Thus, there will be two system models of fundamental interest to us. First there is the continuous-time system dynamics model

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \tag{4-139a}$$

or, more properly,

$$\mathbf{dx}(t) = \mathbf{F}(t)\mathbf{x}(t)\,dt + \mathbf{B}(t)\mathbf{u}(t)\,dt + \mathbf{G}(t)\,\mathbf{d\boldsymbol{\beta}}(t) \tag{4-139b}$$

from which sampled-data measurements are available at times $t_1, t_2, \ldots$ as

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{4-140}$$
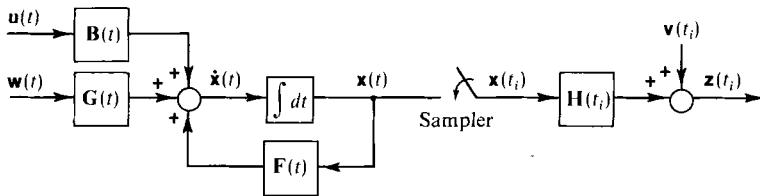
FIG. 4.13 Continuous-time dynamics/discrete-time measurement model.
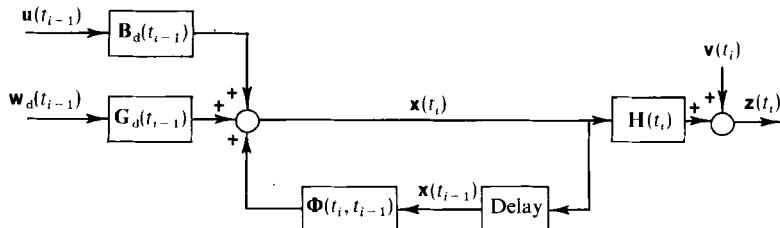
FIG. 4.14 Discrete-time dynamics/discrete-time measurement model.
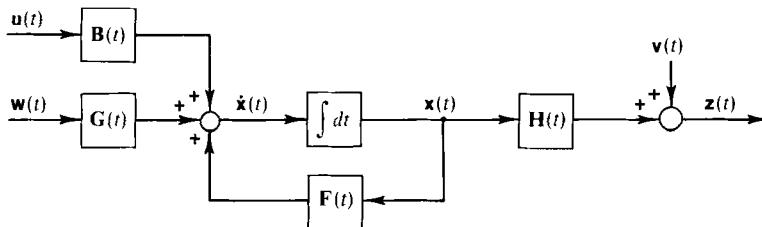
FIG. 4.15 Continuous-time dynamics/continuous-time measurement model.

This model is portrayed schematically in Fig. 4.13. The second model of interest is depicted in Fig. 4.14: a discrete-time dynamics model

$$\mathbf{x}(t_{i+1}) = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \mathbf{B}_d(t_i)\mathbf{u}(t_i) + \mathbf{G}_d(t_i)\mathbf{w}_d(t_i) \qquad (4\text{-}141)$$

with measurements available at discrete times $t_1, t_2, \ldots$, of the same form as in (4-140). Note that if a model is derived originally in the first form, then an "equivalent discrete-time model" can be generated as in Fig. 4.14, but with $\mathbf{G}_d(t_i)$ equal to an n-by-n identity matrix for all times $t_i \in T$.

A third possible model formulation is depicted in Fig. 4.15, consisting of a continuous-time dynamics model (4-139), with measurements continuously available as

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}(t) \qquad (4\text{-}142)$$

with $\mathbf{H}(\cdot)$ an m-by-n matrix of piecewise continuous functions, and $\mathbf{v}(\cdot, \cdot)$ an m-vector-valued continuous-time noise process. This additive noise would be

modeled as a white Gaussian noise with zero mean and covariance kernel

$$E\{\mathbf{v}(t)\mathbf{v}^{\mathrm{T}}(t')\} = \mathbf{R}_c(t)\,\delta(t - t') \qquad (4\text{-}143)$$

It would further be assumed that $\mathbf{v}(t)$ is independent of $\mathbf{x}(t_0)$ and $\mathbf{w}(t')$ or $\boldsymbol{\beta}(t')$ for all t, $t' \in T$ (such an assumption could be relaxed, as mentioned before). Although such a model is of theoretical interest, the sampled-data measurement models are more significant practically, since virtually all estimators and stochastic controllers are implemented on digital computers.

In fact, our attention will be focused upon the continuous-time dynamics/discrete-time measurement model, since this will be the most natural description of the majority of problems of interest. For such a model, the statistics of the system outputs can be calculated explicitly in terms of corresponding state characteristics, which have been described previously. The *mean* of the measurement process at time t_i is

$$\begin{aligned}
\mathbf{m}_z(t_i) = E\{\mathbf{z}(t_i)\} &= E\{\mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i)\} \\
&= \mathbf{H}(t_i)E\{\mathbf{x}(t_i)\} + E\{\mathbf{v}(t_i)\} \\
\mathbf{m}_z(t_i) &= \mathbf{H}(t_i)\mathbf{m}_x(t_i)
\end{aligned} \qquad (4\text{-}144)$$

The output autocorrelation is generated as

$$\begin{aligned}
E\{\mathbf{z}(t_j)\mathbf{z}^{\mathrm{T}}(t_i)\} = {}&E\{\mathbf{H}(t_j)\mathbf{x}(t_j)\mathbf{x}^{\mathrm{T}}(t_i)\mathbf{H}^{\mathrm{T}}(t_i)\} + E\{\mathbf{v}(t_j)\mathbf{v}^{\mathrm{T}}(t_i)\} \\
&+ E\{\mathbf{H}(t_j)\mathbf{x}(t_j)\mathbf{v}^{\mathrm{T}}(t_i)\} + E\{\mathbf{v}(t_j)\mathbf{x}^{\mathrm{T}}(t_i)\mathbf{H}^{\mathrm{T}}(t_i)\}
\end{aligned}$$

But, the third and fourth terms are zero, since we can write

$$\mathbf{x}(t_j) = \boldsymbol{\Phi}(t_j, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t_j} \boldsymbol{\Phi}(t_j, \tau)\mathbf{G}(\tau)\,\mathbf{d}\boldsymbol{\beta}(\tau)$$

or, for the equivalent discrete-time system,

$$\mathbf{x}(t_j) = \boldsymbol{\Phi}(t_j, t_0)\mathbf{x}(t_0) + \sum_{k=1}^{j} \boldsymbol{\Phi}(t_j, t_k)\mathbf{G}_d(t_{k-1})\mathbf{w}_d(t_{k-1})$$

i.e., in either case as the sum of terms, all of which are independent of $\mathbf{v}(t_i)$. Thus, $\mathbf{x}(t_j)$ and $\mathbf{v}(t_i)$ are independent, and so uncorrelated, so that the third term in the previous expression becomes

$$\mathbf{H}(t_j)E\{\mathbf{x}(t_j)\mathbf{v}^{\mathrm{T}}(t_i)\} = \mathbf{H}(t_j)E\{\mathbf{x}(t_j)\}E\{\mathbf{v}^{\mathrm{T}}(t_i)\} = \mathbf{0}$$

since $\mathbf{v}(t_i)$ is assumed zero-mean and similarly for the fourth term. Therefore, the output *autocorrelation* is

$$\boldsymbol{\Psi}_{zz}(t_j, t_i) = E\{\mathbf{z}(t_j)\mathbf{z}^{\mathrm{T}}(t_i)\} = \begin{cases} \mathbf{H}(t_j)E\{\mathbf{x}(t_j)\mathbf{x}^{\mathrm{T}}(t_i)\}\mathbf{H}^{\mathrm{T}}(t_i) & t_j \neq t_i \\ \mathbf{H}(t_i)E\{\mathbf{x}(t_i)\mathbf{x}^{\mathrm{T}}(t_i)\}\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i) & t_j = t_i \end{cases} \qquad (4\text{-}145)$$

Similarly, the *covariance kernel* is

$$\mathbf{P}_{zz}(t_j, t_i) = \begin{cases} \mathbf{H}(t_j)\mathbf{P}_{xx}(t_j, t_i)\mathbf{H}^T(t_i) & t_j \neq t_i \\ \mathbf{H}(t_i)\mathbf{P}_{xx}(t_i, t_i)\mathbf{H}^T(t_i) + \mathbf{R}(t_i) & t_j = t_i \end{cases} \qquad (4\text{-}146)$$

Note that for (4-145) and (4-146), expressions for $E\{\mathbf{x}(t_j)\mathbf{x}^T(t_i)\}$ and $\mathbf{P}_{xx}(t_j, t_i)$ were derived previously for $t_j \geq t_i$ as

$$E\{\mathbf{x}(t_j)\mathbf{x}^T(t_i)\} = \mathbf{\Phi}(t_j, t_i)E\{\mathbf{x}(t_i)\mathbf{x}^T(t_i)\} \qquad \mathbf{P}_{xx}(t_j, t_i) = \mathbf{\Phi}(t_j, t_i)\mathbf{P}_{xx}(t_i)$$
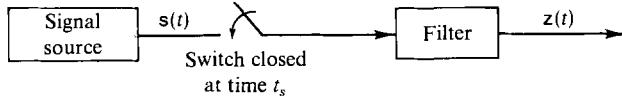
FIG. 4.16   Schematic for Example 4.9.

EXAMPLE 4.9   The following problem, first suggested by Deyst [3], incorporates many of the concepts that have been discussed. Consider Fig. 4.16. A signal source generates a scalar output which can be described as a stationary zero-mean process $s(\cdot, \cdot)$, which is exponentially time-correlated with correlation time T:

$$E\{s(t)\} = 0$$
$$E\{s(t)s(t + \tau)\} = \sigma^2 e^{-|\tau|/T}$$

That is to say, the signal source is started at some time t_0, and the transients are allowed to decay so as to achieve a steady state output process. Then, at time t_s, the switch is closed. It is assumed that at the time just before switch closure, the filter output is zero: $z(t_s^-) = 0$. The filter is a lead-lag type, with a transfer function

$$\frac{z(s)}{s(s)} = \frac{s + a}{s + b}$$

Its amplitude ratio (Bode) plot as a function of frequency is depicted in Fig. 4.17 for the case of $a < b$ ($b < a$ is also possible—this yields a low-pass lead-lag). The objective is to derive an expression for the autocorrelation function of the filter output $z(\cdot, \cdot)$, valid for both time arguments assuming values greater than or equal to t_s.
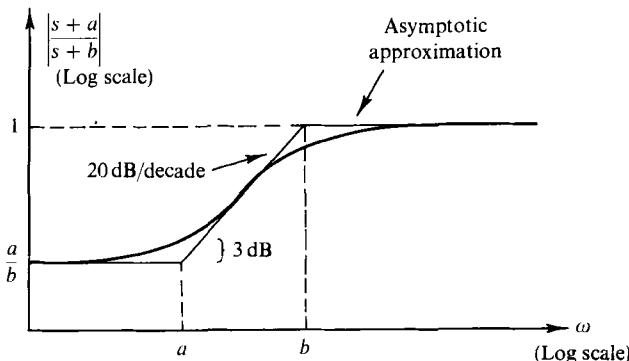
FIG. 4.17   Lead-lag filter amplitude ratio plot.
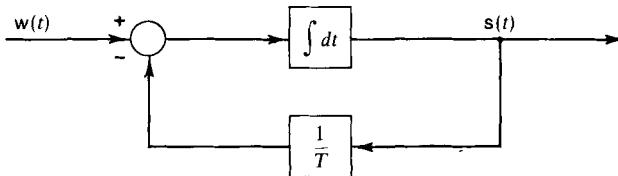
FIG. 4.18   Model of signal generator.

First, we must generate a model of the signal source. A system that generates a process duplicating the characteristics of $s(\cdot,\cdot)$ is depicted in Fig. 4.18: a first order lag driven by zero-mean white Gaussian noise. The output s satisfies the differential equation

$$\dot{s}(t) = -(1/T)s(t) + w(t)$$

where the strength of $w(\cdot,\cdot)$ is as yet unknown: it must be determined so as to yield the appropriate steady state variance of $s(\cdot,\cdot)$. The state transition matrix for this system model is

$$\Phi(t,t') = e^{-(t-t')/T}$$

Therefore, the statistics of $s(\cdot,\cdot)$, assuming $E\{s(t_0)\} = 0$, are given by (4-111) and (4-114) as

$$m_s(t) = E\{s(t)\} = 0$$

$$P_{ss}(t) = E\{s^2(t)\} = e^{-2(t-t_0)/T}E\{s^2(t_0)\} + \int_{t_0}^t e^{-2(t-\tau)/T} Q\, d\tau$$

$$= e^{-2(t-t_0)/T}E\{s^2(t_0)\} + \tfrac{1}{2}QT[1 - e^{-2(t-t_0)/T}]$$

where we are seeking the appropriate value of $Q: E\{w(t)w(t+\tau)\} = Q\,\delta(\tau)$. To obtain stationary characteristics of s, Q must be constant and we must let the transient die out by letting $t_0 \to -\infty$, yielding

$$E\{s^2(t)\} \to \tfrac{1}{2}QT$$

But, from the given autocorrelation·function $\sigma^2 e^{-|\tau|/T}$, this is supposed to equal σ^2, so the desired value of Q is

$$Q = (2/T)\sigma^2$$

Note that the model output obeys (4-118):

$$E\{s(t+\tau)s(t)\} = \Phi(t+\tau,t)E\{s^2(t)\} \to e^{-\tau/T}\sigma^2 \qquad \tau > 0$$

which is the desired form of autocorrelation. In fact, identifying the constant part of the given $\sigma^2 e^{-|\tau|/T}$ as the steady state mean squared value, and the function of τ as the state transition matrix, gives the initial insight that a first order lag is the appropriate system model to propose.

Another procedure for obtaining the appropriate Q would be to seek the steady state solution to (4-120), which for this case becomes the scalar equation

$$\dot{P}_{ss}(t) = -(1/T)P_{ss}(t) - (1/T)P_{ss}(t) + Q = 0$$

Since we desire $P_{ss}(t \to \infty) = \sigma^2$, this again yields $Q = (2/T)\sigma^2$.

For the filter, a state model can be generated as in Example 2.6. Thus, an overall model of the situation is depicted in Fig. 4.19. Note the choice of integrator outputs as state variables x_1 and x_2. Once the switch is closed, the state equations become

$$\dot{x}_1(t) = -(1/T)x_1(t) + w(t), \qquad \dot{x}_2(t) = -bx_2(t) + x_1(t)$$
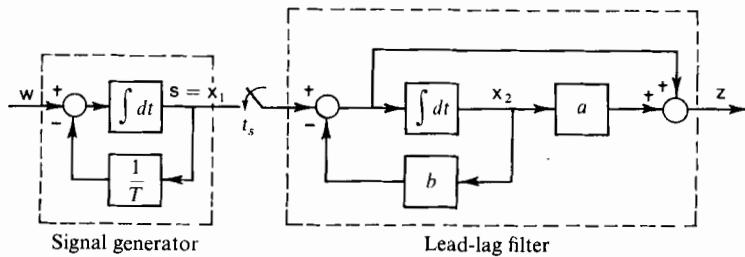
FIG. 4.19   Overall state model for Example 4.9.

and the output equation becomes

$$z(t_i) = ax_2(t_i) + \dot{x}_2(t_i) = ax_2(t_i) - bx_2(t_i) + x_1(t_i) = [a - b]x_2(t_i) + x_1(t_i)$$

In vector notation, this can be written as $\dot{x}(t) = Fx(t) + Gw(t)$, $z(t_i) = Hx(t_i)$:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1/T & 0 \\ 1 & -b \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} w(t)$$

$$z(t_i) = \begin{bmatrix} 1 & (a - b) \end{bmatrix} \begin{bmatrix} x_1(t_i) \\ x_2(t_i) \end{bmatrix}$$

Furthermore, the uncertainties can be described through

$$GQG^T = \begin{bmatrix} 1 \\ 0 \end{bmatrix} Q \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}, \qquad R(t_i) = 0$$

Now (4-145) can be used to write the desired autocorrelation for $t_j \geq t_i$ as

$$E\{z(t_j)z(t_i)\} = HE\{x(t_j)x^T(t_i)\}H^T$$
$$= H\Phi(t_j, t_i)[P_{xx}(t_i) + m_x(t_i)m_x{}^T(t_i)]H^T$$

But, $m_x(t)$ is zero for all time $t \geq t_s$: by time t_s, all transients in $x_1(\cdot)$ have died out, and it is driven by zero-mean noise; $x_2(\cdot)$ starts at $x_2(t_s) = 0$ since $z(t_s{}^-) = 0$, and it is driven by the zero-mean process $x_1(\cdot)$. Thus, the desired result is

$$E\{z(t_j)z(t_i)\} = H\Phi(t_j, t_i)P_{xx}(t_i)H^T$$

and so $\Phi(t_j, t_i)$ and $P_{xx}(t_i)$ must be generated explicitly.

The state transition matrix can be derived using Laplace transforms because of the time-invariant nature of the system:

$$\Phi(s) = [sI - F]^{-1} = \begin{bmatrix} s + (1/T) & 0 \\ -1 & s + b \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{1}{s + (1/T)} & 0 \\ \dfrac{1}{(s + (1/T))(s + b)} & \dfrac{1}{s + b} \end{bmatrix}$$

so that

$$\Phi(t, t') = \begin{bmatrix} e^{-(t-t')/T} & 0 \\ (b - (1/T))^{-1}[e^{-(t-t')/T} - e^{-b(t-t')}] & e^{-b(t-t')} \end{bmatrix}$$

The covariance matrix $\mathbf{P}_{xx}(t_i)$ can be written for any $t_i \geq t_s$ through (4-114) as

$$\mathbf{P}_{xx}(t_i) = \mathbf{\Phi}(t_i, t_s)\mathbf{P}_{xx}(t_s)\mathbf{\Phi}^\mathrm{T}(t_i, t_s) + \int_{t_s}^{t_i} \mathbf{\Phi}(t_i, \tau) \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \mathbf{\Phi}^\mathrm{T}(t_i, \tau)\, d\tau$$

The initial condition is specified by noting $x_1(t_s) = s(t_s)$, and $x_2(t_s)$ is known exactly:

$$\mathbf{P}_{xx}(t_s) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}$$

Substituting these into the expression for $E\{z(t_j)z(t_i)\}$ yields the final result as

$$\begin{aligned}
E\{z(t_j)z(t_i)\} = {}& \sigma^2 \big[c_1{}^2 \exp[-(t_i + t_j - 2t_s)/T] + c_2{}^2 \exp[-b(t_i + t_j - 2t_s)] \\
& - c_1 c_2 (\exp[-b(t_j - t_s) - (t_i - t_s)/T] + \exp[-(t_j - t_s)/T - b(t_i - t_s)]) \big] \\
& + (2\sigma^2/T) \int_{t_s}^{t_i} \big[c_1{}^2 \exp[-(t_i + t_j - 2\tau)/T] + c_2{}^2 \exp[-b(t_i + t_j - 2\tau)] \\
& - c_1 c_2 (\exp[-b(t_j - \tau) - (t_i - \tau)/T] + \exp[-(t_j - \tau)/T - b(t_i - \tau)]) \big]\, d\tau
\end{aligned}$$

where

$$c_1 = \frac{a + b - (2/T)}{b - (1/T)}, \qquad c_2 = \frac{a - (1/T)}{b - (1/T)}$$

Note that the result is a *nonstationary* autocorrelation. To obtain a stationary output, we require not only a time-invariant system driven only by stationary inputs, but also must consider the output only after sufficient time for transients to die out. This stationary result as $t_s \to -\infty$ can also be obtained by convolutions, but the initial transient after time t_s *cannot* be generated through such an approach. ∎

4.11 SHAPING FILTERS AND STATE AUGMENTATION

In many instances, the use of white Gaussian noise models to describe all noises in a real system may not be adequate. It would be desirable to be able to generate empirical autocorrelation or power spectral density data, and then to develop a mathematical model that would produce an output with duplicate characteristics. If observed data were in fact samples from a Brownian motion or stationary Gaussian process with a known rational power spectral density (or corresponding known autocorrelation or covariance kernel), then a linear time-invariant system, or *shaping filter*, driven by stationary white Gaussian noise, provides such a model. If the power spectral density is not rational, it can be approximated as closely as desired by a rational model, and the same procedure followed. Furthermore, if all one knows are the first and second order statistics of a wide-sense stationary process (which is often the case), then a *Gaussian* process with the same first and second order statistics can always be generated via a shaping filter. Time-varying shaping filters are also possible, but we will focus mostly upon models for stationary processes.

The previous example in fact demonstrated the use of such a shaping filter, a first order lag, to duplicate the observed process $s(\cdot, \cdot)$. This section will formalize and extend that development.

Suppose that a system of interest is described by

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{n}(t) \tag{4-147a}$$

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}(t) \tag{4-147b}$$

where $\mathbf{n}(\cdot,\cdot)$ is a nonwhite, i.e., time-correlated, Gaussian noise. Also, suppose that the noise $\mathbf{n}(\cdot,\cdot)$ can be generated by a linear shaping filter:

$$\dot{\mathbf{x}}_f(t) = \mathbf{F}_f(t)\mathbf{x}_f(t) + \mathbf{G}_f(t)\mathbf{w}(t) \tag{4-148a}$$

$$\mathbf{n}(t) = \mathbf{H}_f(t)\mathbf{x}_f(t) \tag{4-148b}$$

where the subscript f denotes filter, and $\mathbf{w}(\cdot,\cdot)$ is a *white* Gaussian noise process. Then the filter output in (4-148b) can be used to drive the system, as shown in Fig. 4.20. Now define the *augmented state vector* process $\mathbf{x}_a(\cdot,\cdot)$ through

$$\mathbf{x}_a(\cdot,\cdot) = \begin{bmatrix} \mathbf{x}(\cdot,\cdot) \\ \mathbf{x}_f(\cdot,\cdot) \end{bmatrix} \tag{4-149}$$

to write (4-147) and (4-148) as an augmented state equation

$$\begin{bmatrix} \dot{\mathbf{x}}(t) \\ \hline \dot{\mathbf{x}}_f(t) \end{bmatrix} = \begin{bmatrix} \mathbf{F}(t) & \vdots & \mathbf{G}(t)\mathbf{H}_f(t) \\ \hline \mathbf{0} & \vdots & \mathbf{F}_f(t) \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \hline \mathbf{x}_f(t) \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \hline \mathbf{G}_f(t) \end{bmatrix} \mathbf{w}(t) \tag{4-150a}$$

$$\dot{\mathbf{x}}_a(t) = \mathbf{F}_a(t) \quad \mathbf{x}_a(t) + \mathbf{G}_a(t) \mathbf{w}(t) \tag{4-150b}$$

and associated output equation

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{H}(t) & \vdots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \hline \mathbf{x}_f(t) \end{bmatrix} + \mathbf{v}(t) \tag{4-150c}$$

$$\mathbf{z}(t) = \mathbf{H}_a(t) \quad \mathbf{x}_a(t) + \mathbf{v}(t) \tag{4-150d}$$

This is again in the form of an overall (augmented) linear system model driven only by *white* Gaussian noise.

An analogous development is possible for the case of time-correlated measurement corruption noise. Let a system of interest be described by

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{w}(t) \tag{4-151a}$$
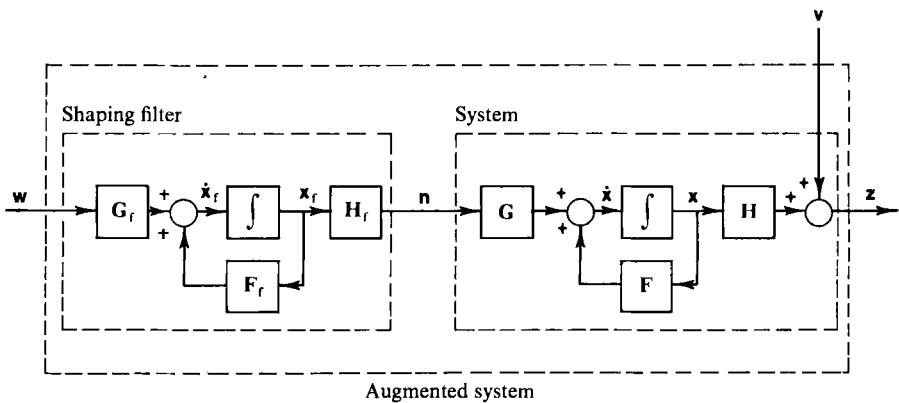
FIG. 4.20 Shaping filter generating dynamic driving noise.

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{n}(t) + \mathbf{v}(t) \tag{4-151b}$$

where $\mathbf{w}(\cdot,\cdot)$ and $\mathbf{v}(\cdot,\cdot)$ are white noises and $\mathbf{n}(\cdot,\cdot)$ is nonwhite. Generate $\mathbf{n}(\cdot,\cdot)$ as the output of a shaping filter driven by white Gaussian $\mathbf{w}_f(\cdot,\cdot)$, as depicted in Fig. 4.21:

$$\dot{\mathbf{x}}_f(t) = \mathbf{F}_f(t)\mathbf{x}_f(t) + \mathbf{G}_f(t)\mathbf{w}_f(t) \tag{4-152a}$$

$$\mathbf{n}(t) = \mathbf{H}_f(t)\mathbf{x}_f(t) \tag{4-152b}$$

The augmented state can be defined as in (4-149) to yield an augmented system description as

$$\begin{bmatrix} \dot{\mathbf{x}}(t) \\ \dot{\mathbf{x}}_f(t) \end{bmatrix} = \begin{bmatrix} \mathbf{F}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_f(t) \end{bmatrix}\begin{bmatrix} \mathbf{x}(t) \\ \mathbf{x}_f(t) \end{bmatrix} + \begin{bmatrix} \mathbf{G}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_f(t) \end{bmatrix}\begin{bmatrix} \mathbf{w}(t) \\ \mathbf{w}_f(t) \end{bmatrix} \tag{4-153a}$$

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{H}(t) & \mathbf{H}_f(t) \end{bmatrix}\begin{bmatrix} \mathbf{x}(t) \\ \mathbf{x}_f(t) \end{bmatrix} + \mathbf{v}(t) \tag{4-153b}$$

which is again in the form of a linear system model driven only by white Gaussian noises. Obvious extensions can allow time-correlated components of both the dynamic driving noise and the measurement corruption noise for a given system.

Certain shaping filter configurations are recurrent and useful enough for process modeling to be discussed individually. These are depicted in Fig. 4.22.
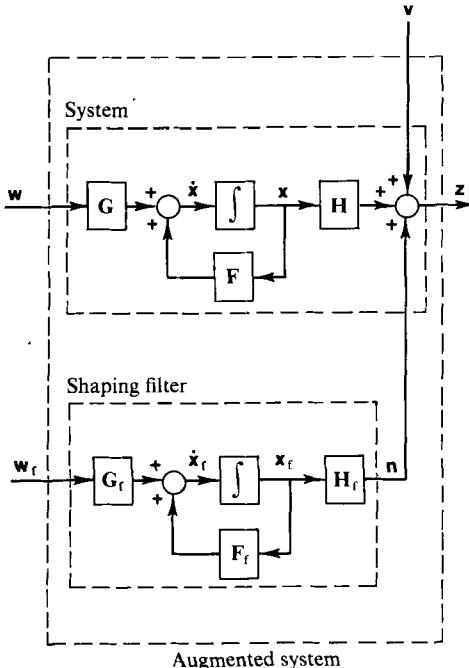
FIG. 4.21   Shaping filter generating measurement corruption noise.
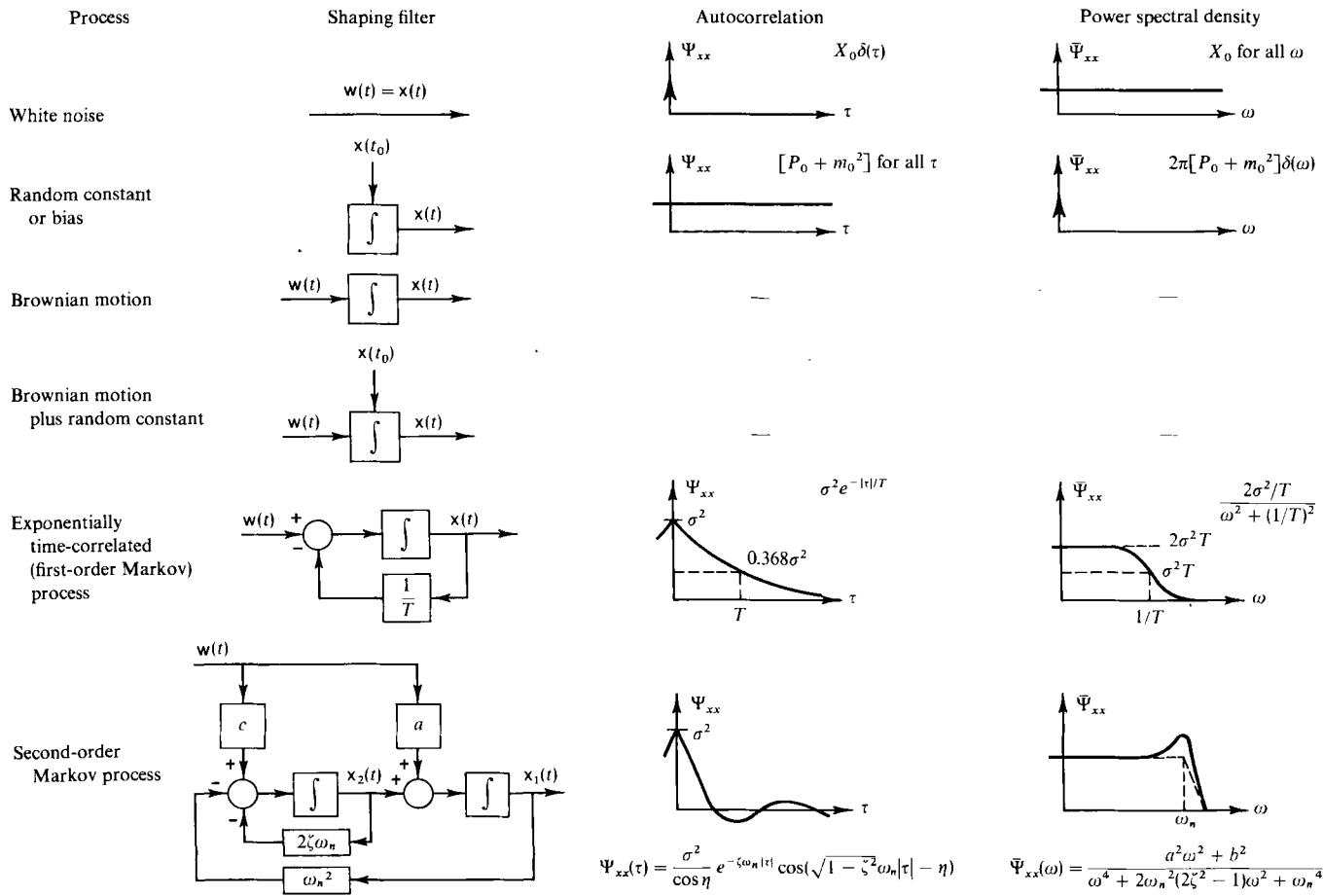
FIG. 4.22 Useful shaping filters.

The trivial case is stationary *white Gaussian noise* itself, of mean m_0 and auto-correlation

$$E\{w(t)w(t + \tau)\} = [P_0 + m_0{}^2]\,\delta(\tau) = X_0\,\delta(\tau) \qquad (4\text{-}154)$$

Second, there is the *random constant* or *bias* model, generated as the output of an integrator with no input, but with an initial condition modeled as a Gaussian random variable $x(t_0)$ with specified mean m_0 and variance P_0. Thus, the defining relationship is

$$\dot{x}(t) = 0 \qquad (4\text{-}155\text{a})$$

starting from the given initial condition $x(t_0)$. Note that this is a degenerate form of shaping filter, in that no noise drives the filter state equation. Since the samples are constant in time, the autocorrelation is constant over all τ, resulting in an impulse power spectral density (all power concentrated at $\omega = 0$):

$$\Psi_{xx}(\tau) = E\{x(t)x(t + \tau)\} = [P_0 + m_0{}^2] \qquad (4\text{-}155\text{b})$$

$$\bar{\Psi}_{xx}(\omega) = 2\pi[P_0 + m_0{}^2]\,\delta(\omega) \qquad (4\text{-}155\text{c})$$

This is a good model for such phenomena as turnon-to-turnon nonrepeatability biases of rate gyros and other sensors: from one period of operation to another, the bias level can change, but it remains constant while the instrument is turned on. Care must be exercised in using such a model in developing an optimal estimator. This model indicates that although you do not know the bias magnitude a priori, you *know* that it does not change value in time. As a result, an optimal filter will estimate its magnitude using initial data, but will essentially disregard all measurements that come later. If it is desired to maintain a viable estimate of a bias that *may* vary slowly (or unexpectedly, as due to instrument failure or degradation), the following shaping filter is a more appropriate bias model.

Brownian motion (random walk) is the output of an integrator driven by white Gaussian noise (heuristically, in view of the development of the previous sections):

$$\dot{x}(t) = w(t); \qquad x(t_0) \triangleq 0 \qquad (4\text{-}156)$$

where $w(\cdot, \cdot)$ has mean zero and $E\{w(t)w(t + \tau)\} = Q\,\delta(\tau)$. The mean equation would be the same as for the random constant, $\dot{m}_x(t) = 0$ or $m_x(t) = m_x(t_0)$, but the second order statistics are different: $\dot{P}_{xx}(t) = Q$ instead of $\dot{P}_{xx}(t) = 0$, so that the mean squared value grows linearly with time, $E\{x^2(t)\} = Q[t - t_0]$. The random walk and random constant can both be represented by the use of only one state variable, essentially just adding the capability of generalizing a random walk to the case of nonzero mean or nonzero initial variance.

Exponentially time-correlated (first order Markov) process models are first order lags driven by zero-mean white Gaussian noise of strength Q. As shown

in Example 4.9, to produce an output with autocorrelation

$$\Psi_{xx}(\tau) = E\{x(t)x(t + \tau)\} = \sigma^2 e^{-|\tau|/T} \qquad (4\text{-}157a)$$

i.e., of correlation time T and mean squared value σ^2 (and mean zero), the model is described by

$$\dot{x}(t) = -(1/T)x(t) + w(t) \qquad (4\text{-}157b)$$

where $Q = 2\sigma^2/T$, or, in other words, $E\{x^2(t)\} = QT/2$. The associated power spectral density is

$$\Psi_{xx}(\omega) = \frac{2\sigma^2/T}{\omega^2 + (1/T)^2} \qquad (4\text{-}157c)$$

This is an especially useful shaping filter, providing an adequate approximation to a wide variety of empirically observed band-limited (wide or narrow band) noises.

A *second order Markov process* provides a good model of oscillatory random phenomena, such as vibration, bending, and fuel slosh in aerospace vehicles. The general form of autocorrelation is

$$\Psi_{xx}(\tau) = E\{x(t)x(t + \tau)\} = \frac{\sigma^2}{\cos\eta} e^{-\zeta\omega_n|\tau|} \cos(\sqrt{1 - \zeta^2}\omega_n|\tau| - \eta) \quad (4\text{-}158a)$$

as depicted in Fig. 4.22 [note that the autocorrelation periodically is negative, i.e., if you knew the value of $x(t)$, then you expect $x(t + \tau)$ to be of opposite sign for those values of time difference τ]. This can be generated by passing a stationary white Gaussian noise $w(\cdot,\cdot)$ of strength $Q = 1$ through a second order system, having a transfer function most generally expressed as

$$G(s) = \frac{as + b}{s^2 + 2\zeta\omega_n s + \omega_n^2} \qquad (4\text{-}158b)$$

or a state description as depicted in Fig. 4.22:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\zeta\omega_n \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} a \\ c \end{bmatrix} w(t) \qquad (4\text{-}158c)$$

where $x_1(t)$ is the system output and

$$a = [(2\sigma^2/\cos\eta)\omega_n \sin(\alpha - \eta)]^{1/2}, \qquad b = [(2\sigma^2/\cos\eta)\omega_n^3 \sin(\alpha + \eta)]^{1/2},$$
$$c = b - 2a\zeta\omega_n, \qquad\qquad \alpha = \tan^{-1}[\zeta/\sqrt{1 - \zeta^2}].$$

In practice, σ^2, η, ζ, and ω_n are chosen to fit empirical data. (See Problem 4.21.) For instance, ζ is chosen to fit the observed resonant peak in power spectral density, from the condition of no peak ($\zeta \in (0.707,1]$) to extreme resonance ($\zeta \ll 1$). The extreme case of $\zeta = 1$ and $a = 0$ is not periodic at all, but provides a steeper rolloff of power spectral density than a first order Markov model.

4.12 POWER SPECTRUM CONCEPTS
AND SHAPING FILTERS

It is useful to be able to express the power spectral density of a wide-sense stationary output of a system directly in terms of the power spectral density of the input and the description of the system itself. Inherent in such a concept are the facts that the input is stationary, the system is time invariant, and we are interested only in a steady state description of the output (i.e., $t_0 \to -\infty$). Not only would such a relationship indicate the effect of a given system on the statistics of a signal, it also will yield an expedient means of synthesizing a shaping filter to duplicate a desired power spectral density from a white noise input. This section develops the desired relationship and then directly applies it to shaping filter design.

The input–output relationship of a single input–single output linear system can be described through a time domain analog of a transfer function, called an impulse response function $G_t(\cdot, \cdot)$, where $G_t(t, t')$ is the system output response at time t due to a unit impulse input applied at time t'. In terms of this function, the output $z(t)$ can be expressed in terms of the input $n(t')$ for $t' \in (-\infty, t]$ as

$$z(t) = \int_{-\infty}^{t} G_t(t, t')n(t')\,dt' \tag{4-159}$$

For any physically realizable system, $G_t(t, t') \equiv 0$ for $t < t'$: the system does not respond to an input before it arrives (the system is "nonanticipative"). If the system is time invariant, then the impulse response function is a function only of the time difference $(t - t')$, and not of t and t' separately, denoted by $G_t(t, t') \triangleq G_t(t - t')$, so that (4-159) becomes

$$z(t) = \int_{-\infty}^{t} G_t(t - t')n(t')\,dt' \tag{4-160a}$$

By defining a change of variables, $(t - t') = \tau$ (so that $dt' = -d\tau$), this becomes

$$z(t) = \int_{0}^{\infty} G_t(\tau)n(t - \tau)\,d\tau \tag{4-160b}$$

This is an ordinary convolution integral relation, the Laplace transform of which yields the multiplicative transfer function relation of Eq. (2-2). Note that physical realizability requires $G_t(\tau) = 0$ for $\tau < 0$ in this time-invariant case.

Now (4-160b) can be used to write the stationary statistics of the output process $z(\cdot, \cdot)$ in terms of those of the input process $n(\cdot, \cdot)$:

$$E\{z(t)\} = E\{n(t)\} \int_{0}^{\infty} G_t(\tau)\,d\tau = \text{const} \tag{4-161a}$$

$$\Psi_{zz}(\tau) = \int_{0}^{\infty} \int_{0}^{\infty} G_t(\tau_1)G_t(\tau_2)\Psi_{nn}(\tau + \tau_1 - \tau_2)\,d\tau_1\,d\tau_2 \tag{4-161b}$$

$$E\{z^2(t)\} = \Psi_{zz}(0) = \int_{0}^{\infty} \int_{0}^{\infty} G_t(\tau_1)G_t(\tau_2)\Psi_{nn}(\tau_1 - \tau_2)\,d\tau_1\,d\tau_2 \tag{4-161c}$$

$$\Psi_{nz}(\tau) = \int_{0}^{\infty} G_t(\tau_1)\Psi_{nn}(\tau - \tau_1)\,d\tau_1 \tag{4-161d}$$

Define $G(\cdot)$ to be the Fourier transform of $G_t(\cdot)$:

$$G(\omega) = \int_{-\infty}^{\infty} G_t(\tau)e^{-j\omega\tau}\,d\tau$$

By considering the Fourier transform of (4-161), the convolutions become product relations.

First, due to physical realizability, $\int_0^{\infty} G_t(\tau)\,d\tau = \int_{-\infty}^{\infty} G_t(\tau)\,d\tau$ in (4-161a). This is then recognized as the Fourier transform evaluated at $\omega = 0$:

$$E\{z(t)\} = E\{n(t)\}G(0) = \text{const} \qquad (4\text{-}162a)$$

The Fourier transform of (4-161b) is

$$\Psi_{zz}(\omega) = \int_{-\infty}^{\infty} \Psi_{zz}(\tau)e^{-j\omega\tau}\,d\tau$$

$$= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} G_t(\tau_1)G_t(\tau_2)\Psi_{nn}(\tau + \tau_1 - \tau_2)\,d\tau_1\,d\tau_2\,e^{-j\omega\tau}\,d\tau$$

But the orders of integration can be changed since $\Psi_{nn}(\cdot)$ is assumed Fourier transformable, so

$$\Psi_{zz}(\omega) = \int_0^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty} G_t(\tau_1)G_t(\tau_2)\Psi_{nn}(\tau + \tau_1 - \tau_2)e^{-j\omega\tau}\,d\tau\,d\tau_2\,d\tau_1$$

$$= \int_0^{\infty} G_t(\tau_1)e^{j\omega\tau_1}\,d\tau_1 \int_0^{\infty} G_t(\tau_2)e^{-j\omega\tau_2}\,d\tau_2$$

$$\times \int_{-\infty}^{\infty} \Psi_{nn}(\tau + \tau_1 - \tau_2)e^{-j\omega(\tau + \tau_1 - \tau_2)}\,d\tau$$

$$= G(-\omega)G(\omega)\Psi_{nn}(\omega)$$

where use has been made of the fact that $G_t(\tau) = 0$ for $\tau < 0$. Remembering that $G(-\omega) = G^*(\omega)$, the final result is

$$\Psi_{zz}(\omega) = G(\omega)G(-\omega)\Psi_{nn}(\omega) = |G(\omega)|^2\Psi_{nn}(\omega) \qquad (4\text{-}162b)$$

Note that this depends only on the magnitude, and not the phase, of $G(\omega)$.

Similarly, (4-161c) and (4-161d) become

$$E\{z^2(t)\} = \text{const} = \Psi_{zz}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{zz}(\omega)\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)G(-\omega)\Psi_{nn}(\omega)\,d\omega \qquad (4\text{-}162c)$$

$$\Psi_{nz}(\omega) = G(\omega)\Psi_{nn}(\omega) \qquad (4\text{-}162d)$$

In much analytical work with linear systems, control engineers use the Laplace transform instead of the Fourier transform. Rather than considering two-sided Laplace transforms, we will treat these as Fourier transforms with a change of variable, replacing ω by s/j, i.e., by letting $s = j\omega$. Since power spectral densities are even functions of ω, there are no odd powers in ω, so $\Psi_{zz}(s)$ is always a *real* function of s: a rational $\Psi_{zz}(s)$ can be obtained from a

rational $\bar{\Psi}_{zz}(\omega)$ by replacing all powers of ω^2 by $(s/j)^2 = -s^2$. The results of (4-162) can then be written directly in terms of functions of s instead of ω, with the only structural change being

$$E\{z^2(t)\} = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} G(s)G(-s)\bar{\Psi}_{nn}(s)\, ds \qquad (4\text{-}162c')$$

Now we can consider shaping filter design: suppose we have a stationary stochastic process $n(\cdot,\cdot)$ with known power spectral density $\bar{\Psi}_{nn}(\cdot)$ (without loss of generality, a white process), and we want to generate from it a process $z(\cdot,\cdot)$ with a desired $\bar{\Psi}_{zz}(\cdot)$. If both spectra are rational (the ratio of polynomials in ω), then the design of the linear time-invariant shaping filter is straightforward, once spectral factorization is understood.

If $\bar{\Psi}_{zz}(\cdot)$ is rational, then it can be written, since it is also even, as

$$\bar{\Psi}_{zz}(\omega) = \frac{a_0 + a_1\omega^2 + a_2\omega^4 + a_3\omega^6 + \cdots}{b_0 + b_1\omega^2 + b_2\omega^4 + b_3\omega^6 + \cdots} \qquad (4\text{-}163a)$$

or

$$\bar{\Psi}_{zz}(s) = \frac{a_0 - a_1 s^2 + a_2 s^4 - a_3 s^6 + \cdots}{b_0 - b_1 s^2 + b_2 s^4 - b_3 s^6 + \cdots} \qquad (4\text{-}163b)$$

$\bar{\Psi}_{zz}(s)$ can then always be factored into the form

$$\bar{\Psi}_{zz}(s) = K\frac{(c_1 - s^2)(c_2 - s^2)\cdots}{(d_1 - s^2)(d_2 - s^2)\cdots} \qquad (4\text{-}163c)$$

Since the coefficients a_i and b_j are all real numbers, the c_i's and d_j's must either be real or occur in complex conjugate pairs. For real positive d_i, the poles are at $\pm\sqrt{d_i}$, as shown in Fig. 4.23a. For complex d_i, the poles are at $\pm\sqrt{d_i} = \pm(e + jf)$, as in Fig. 4.23b. In such a case, there is another d_j that is the complex conjugate of d_i, $d_j = d_i^*$, with roots at $\pm\sqrt{d_j} = \pm\sqrt{d_i^*} = \pm(e - jf)$, as in Fig. 4.23c. Thus, complex roots occur in quadruplets, symmetric about both the real and imaginary axes of the s plane, as in Fig. 4-23d. Similarly, any pure imaginary poles, the case for real negative d_i, appear as doubles; this can be viewed as the case above in the limit of zero separation, and is depicted in Fig. 4.23e. Zeros corresponding to the c_i in (4-163c) are treated analogously.

Now collect all factors of $\bar{\Psi}_{zz}(s)$ which define poles or zeros in the left half s plane, and denote the product of all of these factors and $\sqrt{K}$ [recall (4-163c)] as $\bar{\Psi}_{zz}(s)_{\mathrm{L}}$. Correspondingly generate $\bar{\Psi}_{zz}(s)_{\mathrm{R}}$ of right half plane factors and $\sqrt{K}$. Since pure imaginary roots always appear as doubles, one would be associated with $\bar{\Psi}_{zz}(s)_{\mathrm{L}}$ and the other with $\bar{\Psi}_{zz}(s)_{\mathrm{R}}$. This yields the *spectral factorization*:

$$\bar{\Psi}_{zz}(s) = \bar{\Psi}_{zz}(s)_{\mathrm{L}}\bar{\Psi}_{zz}(s)_{\mathrm{R}} \qquad (4\text{-}164)$$

where all poles and zeros of $\bar{\Psi}_{zz}(s)_{\mathrm{L}}$ are in the left half plane, and all of $\bar{\Psi}_{zz}(s)_{\mathrm{R}}$ are in the right half plane. Note that, because of the symmetry about both

Real positive d_i

(a)

Complex d_i

(b)   (c)   (d)

Real negative d_i

(e)
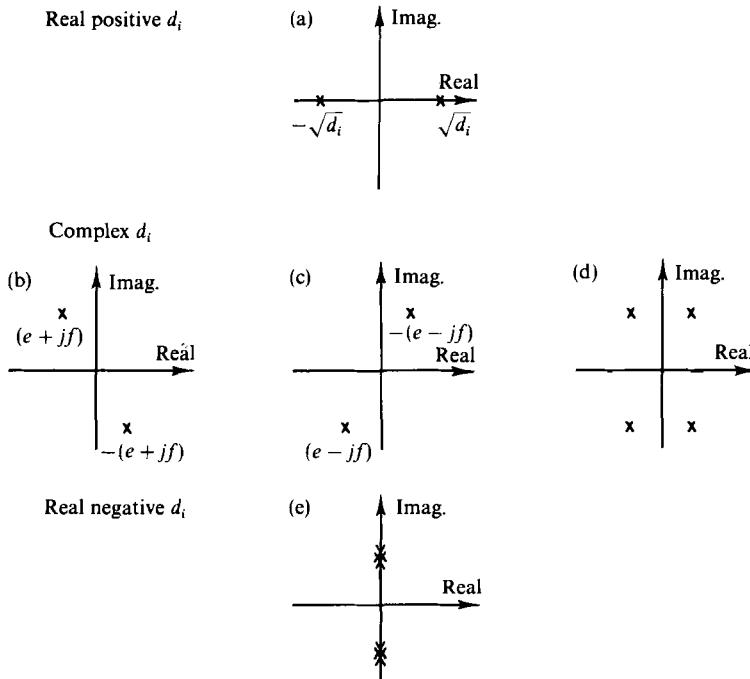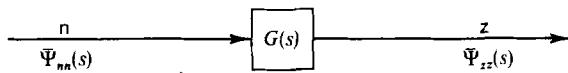
FIG. 4.23 s-Plane plots of roots of rational $\Psi_{zz}(s)$.

FIG. 4.24   Shaping filter design.

coordinate axes,

$$\Psi_{zz}(s)_\mathrm{R} = \Psi_{zz}(-s)_\mathrm{L} \qquad (4\text{-}165)$$

Now we have the situation depicted in Fig. 4.24: we know $\Psi_{nn}(s)$ and $\Psi_{zz}(s)$ for all s, and wish to determine the appropriate $G(s)$ to describe the shaping filter. From (4-162b),

$$\Psi_{zz}(s) = G(s)G(-s)\Psi_{nn}(s)$$

which can be written in factored form as

$$\Psi_{zz}(s)_\mathrm{L}\Psi_{zz}(s)_\mathrm{R} = G(s)G(-s)\Psi_{nn}(s)_\mathrm{L}\Psi_{nn}(s)_\mathrm{R}$$

or, equivalently in view of (4-165),

$$\Psi_{zz}(s)_\mathrm{L}\Psi_{zz}(-s)_\mathrm{L} = G(s)G(-s)\Psi_{nn}(s)_\mathrm{L}\Psi_{nn}(-s)_\mathrm{L}$$

Thus, by letting the shaping filter be described by

$$G(s) = \Psi_{zz}(s)_\mathrm{L}/\Psi_{nn}(s)_\mathrm{L} \qquad (4\text{-}166)$$

it will have all of its poles in the left half plane (and thus be *stable*) and all of its zeros also in the left half plane (and thus be *minimum phase*: for a given amplitude-versus-frequency Bode plot, this form has the least phase lag).

EXAMPLE 4.10   Use a white (Gaussian) noise with $\Psi_{nn}(\omega) = Q$ for all ω to generate an exponentially time-correlated (Gaussian) noise with

$$\Psi_{zz}(\omega) = \frac{2\sigma^2/T}{\omega^2 + (1/T)^2}$$

First replace ω^2 by $(-s^2)$ to generate $\Psi_{zz}(s)$:

$$\Psi_{zz}(s) = \frac{2\sigma^2/T}{(1/T)^2 - s^2} = \frac{2\sigma^2/T}{[(1/T) - s][(1/T) + s]}$$

Perform spectral factorization to obtain

$$\Psi_{zz}(s) = \frac{\sqrt{2/T}\sigma}{(1/T) + s}\frac{\sqrt{2/T}\sigma}{(1/T) - s} = [\Psi_{zz}(s)_L][\Psi_{zz}(s)_R]$$

and similarly $\Psi_{nn}(s)_L = \sqrt{Q}$. The desired shaping filter can then be expressed as

$$G(s) = \frac{\Psi_{zz}(s)_L}{\Psi_{nn}(s)_L} = \frac{\sqrt{2/TQ}\sigma}{s + (1/T)}$$

Note that if we let $n(\cdot, \cdot)$ be a white noise with $Q = 2\sigma^2/T$, then passing it through a first order lag, $G(s) = 1/[s + (1/T)]$, yields the desired result. This agrees with the results found previously by time-domain shaping filter design techniques.   ■

EXAMPLE 4.11   To illustrate the case of poles or zeros on the imaginary axis, consider generating a signal $z(\cdot, \cdot)$ with

$$\Psi_{zz}(\omega) = \frac{a\omega^2}{b^2 + \omega^2}$$

from white noise $n(\cdot, \cdot)$ with $\Psi_{nn}(\omega) = Q$. Replace ω^2 by $(-s^2)$ and perform spectral factorization:

$$\Psi_{zz}(s) = \frac{-as^2}{b^2 - s^2} = \lim_{\varepsilon \to 0} \frac{\varepsilon^2 - as^2}{b^2 - s^2} = \lim_{\varepsilon \to 0} \frac{\varepsilon + \sqrt{a}s}{b + s}\frac{\varepsilon - \sqrt{a}s}{b - s}$$

$$= \left[\frac{\sqrt{a}s}{b + s}\right]\left[\frac{-\sqrt{a}s}{b - s}\right] = [\Psi_{zz}(s)_L][\Psi_{zz}(s)_R]$$

where the use of ε allows proper assignment of signs. The shaping filter can be expressed as

$$G(s) = \frac{\Psi_{zz}(s)_L}{\Psi_{nn}(s)_L} = \frac{(\sqrt{a}s)/(b + s)}{\sqrt{Q}} = \sqrt{\frac{a}{Q}}\frac{s}{s + b}   ■$$

4.13   GENERATING PRACTICAL SYSTEM MODELS

In order to generate a model of the errors and uncertainty in a physical device (sensor, actuator, etc.) or any other type of system, empirical test data is collected. For example, a gyro output signal might be recorded over a period of hours or days when subjected to known (as especially zero) inputs. Or, a number of

gyros might be tested in this manner to obtain "population statistics" on all gyros of that particular type.

Conceptually, sample autocorrelations can be generated by taking one set of data and generating all products of the form $x(t_i)x(t_j)$ for all discrete times t_i and t_j of interest. This would correspond to $x(t_i, \omega_k) \cdot x(t_j, \omega_k)$ for a single $\omega_k \in \Omega$. Then other sets of data would provide similar product values for different values of k. Averaging over N sets of data would yield the approximate *sample auto-correlation* $\tilde{\Psi}_{xx}(t_i, t_j; N)$, where

$$\tilde{\Psi}_{xx}(t_i, t_j; N) \triangleq \frac{1}{N} \sum_{k=1}^{N} x(t_i, \omega_k)x(t_j, \omega_k) \qquad (4\text{-}167)$$

If the process $x(\cdot, \cdot)$ is stationary, then the sample autocorrelation is a function of only the time difference $(t_i - t_j)$, and so fewer products are required to define the function totally: $\tilde{\Psi}_{xx}(t_i - t_j; N)$ could be evaluated by using (4-167) for any single value of t_i. Often the ergodic assumption is made in this case, yielding another *sample autocorrelation* $\Psi'_{xx}(m \Delta t; N)$ for $m = 0, \pm 1, \ldots,$ $\pm(N - 1)$ as the *time* average

$$\Psi'_{xx}(m \Delta t; N) \triangleq \frac{1}{N} \sum_{i=0}^{N-|m|-1} x(t_i, \omega_k)x(t_i + |m| \Delta t, \omega_k) \qquad (4\text{-}168)$$

This uses N samples from a *single* realization of $x(\cdot, \cdot)$, equally spaced at a sample period of Δt, to evaluate an approximate autocorrelation as an average of all possible products of samples separated by $m \Delta t$ seconds [there are few such products for m almost as large as N, and $\Psi'_{xx}(m \Delta t; N)$ is assumed to be zero for $m \geq N$)]. The corresponding *estimate of the power spectral density function* can be defined as the discrete Fourier transform of $\Psi'_{xx}(\cdot; N)$:

$$\bar{\Psi}'_{xx}(\omega; N) = \sum_{m=-\infty}^{\infty} \Psi'_{xx}(m \Delta t; N)e^{-jm\omega\Delta t} \qquad (4\text{-}169)$$

Prior to the advent of fast Fourier transform (FFT) algorithms, (4-168) and (4-169) were used directly for generating approximate autocorrelations and power spectral densities from sampled data signals of the form $x(t_0), x(t_1), \ldots,$ $x(t_{N-1})$. Now a somewhat different sequence of computations is performed to generate the desired sample functions [7, 10]. First, the fast Fourier transform of the *data* is generated for $i = 0, 1, 2, \ldots, N - 1$ as

$$\bar{x}(i\omega_s; N) = \sum_{k=0}^{N-1} x(t_0 + k \Delta t)e^{-jk\Delta t\, i\omega_s}$$

$$= \sum_{k=0}^{N-1} x(t_0 + k \Delta t)(e^{-j2\pi/N})^{ki} \qquad (4\text{-}170)$$

where the frequency spacing ω_s between the discrete Fourier transform values is chosen to be $\omega_s = 2\pi/(N \Delta t)$ to provide adequate frequency spacing according

to the "sampling theorem." Second, a power spectral density estimate is computed from the result of (4-170) and

$$\Psi'_{xx}(i\omega_s; N) = \frac{1}{N} |\bar{x}(i\omega_s; N)|^2 \tag{4-171}$$

Computing this for $i = 0, 1, \ldots, N - 1$ requires $2N$ multiplications, since real and imaginary parts are squared separately. The autocorrelation $\Psi'_{xx}(m \Delta t; N)$ for $m = 0, 1, \ldots, N - 1$ is computed as the inverse fast Fourier transform of (4-171). To smooth the result, decreasing the distortion due to only a finite set of data being used, this $\Psi'_{xx}(m \Delta t; N)$ is multiplied by a "window function," i.e., each of its N values is multiplied by a predetermined coefficient, to yield a better (in terms of bias and variance properties) estimate of the autocorrelation, denoted as $\Psi''_{xx}(m \Delta t; N)$ for $m = 0, 1, \ldots, N - 1$. Finally, a better estimate of power spectral density, $\bar{\Psi}''_{xx}(i\omega_s; N)$, $i = 0, 1, \ldots, N - 1$, is computed as the fast Fourier transform of $\Psi''_{xx}(\cdot; N)$. Sometimes additional processing, as averaging of data subsets and "prewhitening," are employed to produce even better estimates of autocorrelation and/or power spectral density. Although this procedure involves a substantial number of computations, it is significantly faster than the straightforward use of (4-168) and (4-169), due to the capabilities of FFT algorithms.

Once such empirical data is generated, shaping filters can be produced by curve-fitting these sample functions. Any degree of complexity of the filters can be provided, depending on the complexity of the curves used to fit the data. This is illustrated by the following two examples.

EXAMPLE 4.12   Suppose laboratory tests of a gyro yield empirical drift rate autocorrelation and power spectral density data as portrayed in Figs. 4.25a and b. A reasonable fit to the autocorrelation data would be a curve of the form

$$\Psi_{xx}(\tau) = \sigma^2 e^{-|\tau|/T} + B$$

which describes a combination of a random bias and an exponentially time-correlated component. The values of the parameters σ, T, and B can be determined so as to provide the best fit of the assumed model to the data.

Looking at the power spectral density data indicates that an additional wideband component (modeled as white) should also be included [the corresponding narrow pulse of $\Psi_{xx}(\tau)$ at $\tau = 0$ is hard to discern from data]. A reasonable curve-fit here is

$$\bar{\Psi}_{xx}(\omega) = \frac{2\sigma^2/T}{\omega^2 + (1/T)^2} + Q$$

(Note that the bias is difficult to discern accurately from this data.)

This yields a gyro drift model as depicted in Fig. 25c, with defining statistics

$$E\{x_1^2(t_0)\} = \sigma^2, \qquad E\{w_1(t)w_1(t + \tau)\} = [2\sigma^2/T]\delta(\tau)$$
$$E\{x_2^2(t_0)\} = B, \qquad E\{w_2(t)w_2(t + \tau)\} = Q\delta(\tau)$$

Naturally, more complicated fitted curves would yield more complex drift models.   ∎

(a)

$$\sigma^2 e^{-|\tau|/T} + B$$

Data points

B

Ψ_{xx}

τ

(b)

$$\frac{2\sigma^2/T}{\omega^2 + (1/T)^2} + Q$$

Data points

Q

Ψ_{xx}

ω

(c)

$w_1(t)$ + $-$ $\int dt$ $x_1(t)$

$\frac{1}{T}$

$w_2(t)$ $+$ $+$ $x(t)$

i.c. $= x_2(t_0)$ $+$

$\int dt$ $x_2(t)$
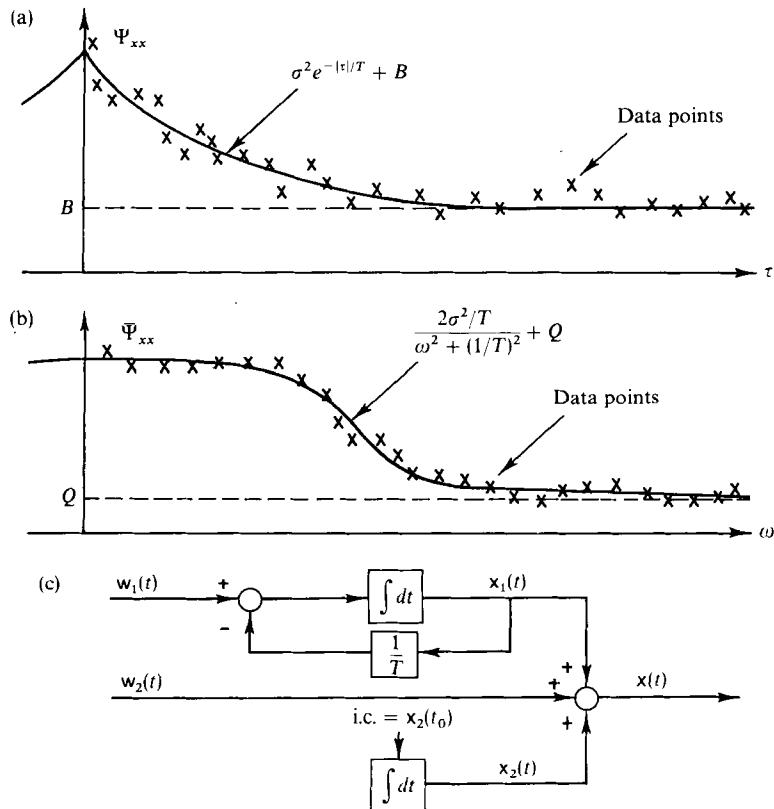
FIG. 4.25 Gyro drift model. (a) Curve-fitting empirical autocorrelation data. (b) Curve-fitting empirical power spectral density data. (c) Shaping filter.

Ψ_{xx}

Data points

$\dfrac{Qb^2}{\omega_n^{\,4}}$

Fitted curve

ω_n ω_p ω
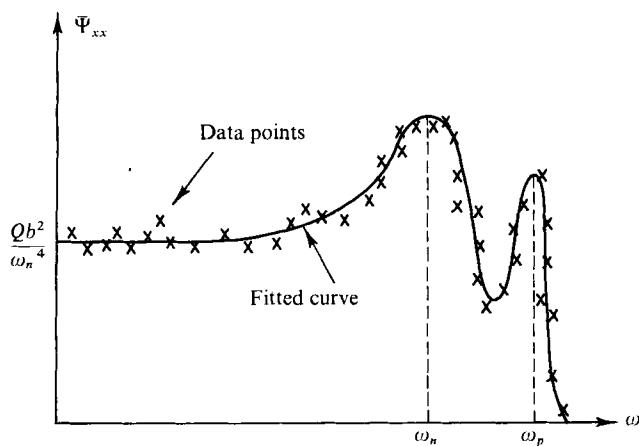
FIG. 4.26 Power spectral density function for Example 4.13.

EXAMPLE 4.13 Suppose vibration testing of a structure produced an acceleration power spectral density at a certain location as in Fig. 4.26. Except for the peaking at $\omega = \omega_p$, the data are well fit by

$$\Psi_{xx}(\omega) = \frac{Qb^2}{\omega^4 + 2\omega_n^2(2\zeta^2 - 1)\omega^2 + \omega_n^4}$$

which is generated by passing white noise of strength Q through a second order shaping filter described through

$$G_1(s) = \frac{b}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

The peak at ω_p can be generated by cascading a "notch filter" of the form

$$G_2(s) = \frac{s^2 + 2\zeta_1\omega_p s + \omega_p^2}{s^2 + 2\zeta_2\omega_p s + \omega_p^2}$$

with the shaping filter described by $G_1(s)$. The size of the resonant peak at ω_p is then adjusted by controlling the magnitude of ζ_1 and ζ_2. Unlike most applications of notch filters, we want to accentuate the signal content in the "notch" region rather than attenuate it, so we require $\zeta_2 < \zeta_1$. ∎

4.14 SUMMARY

Stochastic processes were defined and then characterized through an infinite array of joint distribution functions. A practical, though generally only partial, characterization was then developed in terms of the first two moments: the mean value function and the correlation or covariance kernel. This statistical knowledge was shown to be complete for Gaussian processes, and very readily generated for Gauss–Markov processes. For wide-sense stationary processes, another useful characterization was developed in the form of power spectral density.

A basic system model structure in the form of linear state dynamics driven only by known inputs and white Gaussian noise, with a linear measurement corrupted by additive white Gaussian noise, was motivated and shown to be widely applicable. With such a structure in mind, linear stochastic differential equations (4-121) and their solutions (4-122) were developed properly through stochastic integrals and Brownian motion. The Gauss–Markov state stochastic process could then be characterized by its mean (4-123), covariance (4-114) and (4-120), and covariance kernel (4-118).

With measurements typically available on a sampled-data basis as in (4-136) an overall system model was developed. An equivalent discrete-time system model (4-125) was also developed to describe such an output process from a physical system. Moreover, the output process characteristics were defined in terms of the state process description obtained previously: (4-144)–(4-146) portray the mean, correlation, and covariance kernel for this output process.

Finally, the concept of a shaping filter applied the proposed general model structure to the problem of synthesizing a mathematical model to duplicate the

characteristics of empirically observed processes. Both time-domain and power spectral density techniques were exploited in this synthesis procedure. The generation of empirical autocorrelation and power spectrum data was discussed, and curve-fitting these data then allowed complete definition of the appropriate shaping filter.

At this point, we have adequate linear stochastic models for both static and dynamic systems. These will be exploited extensively in estimation and control and will also provide insights into nonlinear models and associated estimation and control algorithms.

REFERENCES

1. Åström, K. J., "On a First-Order Stochastic Differential Equation," *Internat. J. Control* 1, 301–326 (1965).
2. Davenport, W. B., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*. McGraw-Hill, New York, 1958.
3. Deyst, J. J., "Estimation and Control of Stochastic Processes," unpublished course notes. M.I.T. Department of Aeronautics and Astronautics, Cambridge, Massachusetts, 1970.
4. Doob, J. L., *Stochastic Processes*. Wiley, New York, 1953.
5. Kailath, T., and Frost, P., "Mathematical Modeling of Stochastic Processes," *Stochastic Problems in Control, Proc. Symp. AACC, Ann Arbor, Michigan* (June 1968).
6. Laning, J. H. Jr., and Battin, R. H., *Random Processes in Automatic Control*. McGraw-Hill, New York, 1956.
7. Oppenheim, A. V., and Schafer, R. W., *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1975.
8. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 1965.
9. Parzen, E., *Stochastic Processes*. Holden-Day, San Francisco, California, 1962.
10. Schwartz, M., and Shaw, L., *Signal Processing: Discrete Spectral Analysis, Detection, and Estimation*. McGraw-Hill, New York, 1975.
11. Wax, N., *Selected Papers on Noise and Stochastic Processes*. Dover, New York, 1954.
12. Widnall, W. S., *Applications of Optimal Control Theory to Computer Controller Design*. M.I.T. Press, Cambridge, Massachusetts, 1968.
13. Wong, E., *Stochastic Processes in Information and Dynamical Systems*. McGraw-Hill, New York, 1971.
14. Wong, E., and Zakai, M., "On the Relation Between Ordinary and Stochastic Differential Equations," *Internat. J. Eng. Sci.* 3, 213–229 (1965).

PROBLEMS

4.1 Consider the variable x(t) defined by

$$x(t) = \sum_{i=0}^{N} \Delta x(t_i)$$

where the variables $\Delta x(t_i)$ are independent and Gaussian scalars with statistics

$$E\{\Delta x(t_i)\} = 0, \qquad E\{\Delta x^2(t_i)\} = q[t_{i+1} - t_i]$$

Find the characteristic function, $\phi_x(\mu, t)$, of x(t) and use it to determine the first two moments of x(t).

4.2   Verify the integral below where $\beta(t)$ is Brownian motion:

$$\int_{t_0}^{t} m(\tau)\, d\beta(\tau) = m(t)\beta(t) - m(t_0)\beta(t_0) - \int_{t_0}^{t} \beta(\tau)\, dm(\tau)$$

4.3   Prove that Eq. (4-100) is valid by using fundamental definitions for a finite partitioning of the time axis, and then taking the limit as the time cuts become infinitesimally fine.

4.4   In deriving Eq. (4-145), the following relation was used:

$$\mathbf{x}(t_j) = \mathbf{\Phi}(t_j, t_0)\mathbf{x}(t_0) + \sum_{k=1}^{j} \mathbf{\Phi}(t_j, t_k)\mathbf{G_d}(t_{k-1})\mathbf{w_d}(t_{k-1})$$

Show that this is the solution to the stochastic difference equation (4-133) with $\mathbf{u}(t_i) \equiv \mathbf{0}$ for all time.

4.5   Let $\beta(\cdot, \cdot)$ be a scalar Brownian motion process with statistics

$$E\{\beta(t)\} = 0, \qquad E\{\beta(t)^2\} = t$$

The process $\mathbf{x}(\cdot, \cdot)$ satisfies the stochastic differential equation

$$d\mathbf{x}(t) = \beta(t)\cos t\, dt + \sin t\, d\beta(t)$$

Determine the variance of $\mathbf{x}(t)$.

Explain how you would determine the variance if the equation were, instead,

$$d\mathbf{x}(t) = \mathbf{x}(t)\cos t\, dt + \sin t\, d\beta(t)$$

4.6   Let $\mathbf{x}(\cdot, \cdot)$ be a discrete-time process satisfying

$$\mathbf{x}(t_i) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \mathbf{G_d}(t_{i-1})\mathbf{w_d}(t_{i-1})$$

where $\mathbf{w_d}(\cdot, \cdot)$ is a white Gaussian process with mean $\overline{\mathbf{w}}_d(t_i)$ for all t_i, and covariance kernel $\mathbf{Q_d}(t_i)\delta_{ij}$. Show that $\mathbf{x}(\cdot, \cdot)$ can also be generated by

$$\mathbf{x}(t_i) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \mathbf{G_d}(t_{i-1})\mathbf{u}(t_{i-1}) + \mathbf{G_d}(t_{i-1})\mathbf{w_d}'(t_{i-1})$$

where the deterministic input $\mathbf{u}(t_i) \triangleq \overline{\mathbf{w}}_d(t_i)$ for all t_i and $\mathbf{w_d}'(\cdot, \cdot)$ is a zero-mean white Gaussian noise with covariance kernel $\mathbf{Q_d}(t_i)\delta_{ij}$.

4.7   An engineer had the task of simplifying a system model for eventual filter implementation. He was given a basic model of a noise input $n(\cdot, \cdot)$ into a physical system as in Fig. 4.P1a, where $w_1(\cdot, \cdot)$ and $w_2(\cdot, \cdot)$ are independent white Gaussian noises, each zero-mean and of unit strength (variance kernel of one times the delta function). He has proposed that the model depicted in Fig. 4.P1b is equivalent to the original, and it requires one less noise source in the overall system model. Do you agree that this is equivalent? Explain.

4.8   Suppose you are investigating the system modeled as in Fig. 4.P2. The noises n_1, n_2, n_3, and n_4 are zero-mean white Gaussian noises independent of x_1 and x_2 histories with variances

$$E\{n_i(t)n_i(t + \tau)\} = N_i\,\delta(\tau)$$

for $i = 1, 2, 3, 4$ and N_1, N_2, N_3, and N_4 specified values,

$$E\{n_2(t)n_3(t + \tau)\} = K_{23}\,\delta(\tau)$$

and other cross-correlations zero.

The two sampling devices are corrupted by zero-mean white Gaussian noise sequences d_1 and d_2 with

$$E\{d_1(t_i)d_1(t_j)\} = D_1\,\delta_{ij}, \qquad E\{d_2(t_i)d_2(t_j)\} = D_2\,\delta_{ij}, \qquad E\{d_1(t_i)d_2(t_j)\} = D_3\,\delta_{ij}$$

and the d_i's are independent of the x_i's and n_i's.
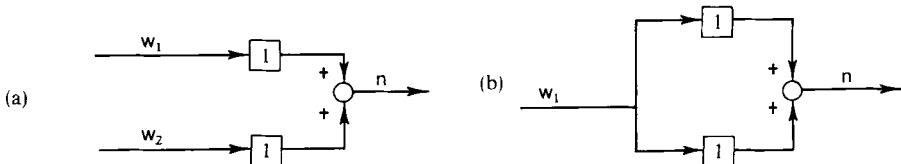
FIG. 4.P1
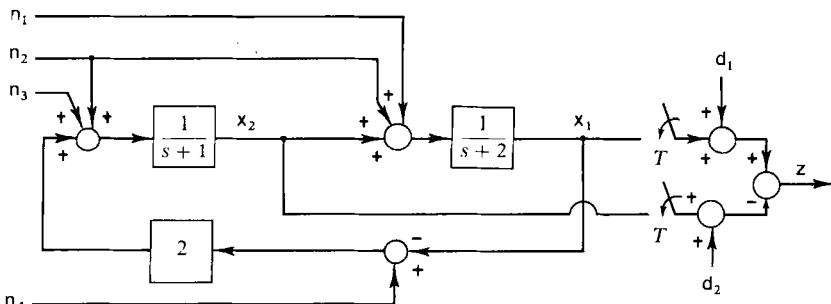
FIG. 4.P2
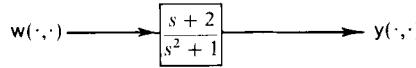
FIG. 4.P3

Develop the linear state variable stochastic differential equations to describe the system, using x_1 and x_2 as states.

Obtain the differential equations for the elements of the covariance matrix to describe the evolution of the second moment of $\mathbf{x} = [x_1 x_2]^T$. Assuming that the solution to the differential equation is available, can you generate an expression for the time-varying autocorrelation function of the output, $E\{z(t_i)z(t_j)\}$ in terms of the elements of this covariance matrix solution?

Can you generate an equivalent system description with fewer noise inputs than the six originally defined? *Specifically* describe such a system model and the statistics of the noises used to replace the original six.

4.9 The transfer function model for a system is given as in Fig. 4.P3, where $w(\cdot,\cdot)$ is a white Gaussian noise with

$$E\{w(t)\} = 0, \qquad E\{w(t)w(t + \tau)\} = \delta(\tau)$$

If the system starts at rest at $t = 0$, determine the variances of $y(t)$ and $\dot{y}(t)$ for $t \geq 0$. (Modeling suggestion: Standard controllable form is convenient.)

4.10 A random process with power spectral density

$$\Psi_{xx}(\omega) = A/(a^2 + \omega^2)$$

drives a first order lag, described by

$$T(s) = 1/(1 + \tau s)$$

What is the steady state mean squared value of y(t), the output of the lag? Note that an expression of the form

$$E\{y^2\} = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \Psi_{yy}(s)\, ds$$

can be evaluated by the Cauchy residue theorem as

$$E\{y^2\} = \sum \{\text{residues of } \Psi_{yy}(s) \text{ in left half plane}\}$$

The result of such calculations is often tabulated in control texts and handbooks.

4.11   A lead-lag network is driven by a scalar white Gaussian noise process w($\cdot$, $\cdot$). The network transfer function is

$$x(s)/w(s) = (s + a)/(s + b)$$

Statistics of the input w($\cdot$, $\cdot$) are

$$E\{w(t)\} = 0, \qquad E\{w(t)w(t + \tau)\} = \delta(\tau)$$

All initial conditions are zero at $t = 0$. Find the nonstationary autocorrelation function for the output x($\cdot$, $\cdot$), $E\{x(t_2)x(t_1)\}$ for all t_1 and t_2, $0 \le t_1 \le t_2$. Note that

$$\int_{t_1}^{t_2} A(\tau)\delta(\tau - t_1)\, d\tau = \tfrac{1}{2}A(t_1)$$

for $A(\cdot)$ a general time function, and the t_1 in the argument of the delta function equal to the lower limit of integration.

4.12   Given the stochastic *vector* differential equation

$$d\mathbf{x}(t) = \mathbf{x}(t)\, dt + d\boldsymbol{\beta}(t)$$

with the initial conditions

$$E\{\mathbf{x}(t_0)\} = \mathbf{0}, \qquad E\{\mathbf{x}(t_0)\mathbf{x}^{\mathrm{T}}(t_0)\} = \mathbf{P}_0$$

and where $\boldsymbol{\beta}(\cdot, \cdot)$ is a vector Brownian motion with

$$E\{\boldsymbol{\beta}(t)\} = \mathbf{0}, \qquad E\{[\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t')][\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t')]^{\mathrm{T}}\} = \mathbf{Q}|t - t'|$$

Determine $E\{\mathbf{x}(t_2)\mathbf{x}^{\mathrm{T}}(t_1)\}$ for $t_0 < t_1 < t_2$.

4.13   Consider the discrete-time process x($\cdot$, $\cdot$) defined by

$$x(i + 1) = [(i + 1)/(i + 2)]x(i)$$

with x(0) a Gaussian random variable with mean $\hat{x}_0$ and variance P_0. Determine the mean and mean square functions, and variance and correlation kernels for the process. Repeat this for

$$x(i + 1) = [(i + 1)/(i + 2)]x(i) + w_d(i)$$

where $w_d(i)$ is zero-mean white Gaussian noise of strength Q_d for all i.

4.14   A stationary process x($\cdot$, $\cdot$) with zero mean and autocorrelation $e^{-\alpha|\tau|}$ is applied at $t = 0$ to a linear system with impulse response $h(t) = e^{-\beta t}U(t)$, where $U(t)$ is the unit step. Find the autocorrelation $P_{yy}(t_1, t_2)$ of the resulting output y($\cdot$, $\cdot$).

Also generate the mean squared value of y(t) for all t. Use both spectral analysis and state space analysis (generating x($\cdot$, $\cdot$) as the output of a shaping filter) to solve this problem.

4.15 Now suppose that the process $x(\cdot, \cdot)$ of the previous problem is applied instead to an integrator starting at rest at $t = 0$. Show that the variance of the output of the integrator is

$$P_{yy}(t) = (2/\alpha^2)(\alpha t - 1 + e^{-\alpha t})$$

and thus never converges to a stationary process as $t \to \infty$.

4.16 Einstein in 1905 gave a solution to the Brownian motion problem. He assumed the visible particles were large compared to the mean free path of the molecules of the fluid, so that the equations of motion of a visible particle would be well approximated by

$$\dot{x}(t) = v(t), \qquad m\dot{v}(t) = -cv(t) + f(t)$$

where $x(t) =$ position of particle, $v(t) =$ velocity of particle, $m =$ mass of particle, $c =$ Stokes's viscous force coefficient (constant), and $f(t) =$ random force due to collision with molecules. The mean time between collisions is very short and $f(t)$ is well approximated by Gaussian white noise, with

$$E\{f(t)\} = 0, \qquad E\{f(t)f(t + \tau)\} = q\,\delta(\tau), \qquad q = \text{const}$$

Assuming that

$$E\{x(0)\} = E\{v(0)\} = E\{x(0)^2\} = E\{v(0)^2\} = E\{x(0)v(0)\} = 0$$

determine expressions for

$$E\{v(t)^2\}, \quad E\{v(t)x(t)\}, \quad \text{and} \quad E\{x(t)^2\}$$

4.17 Recall Problem 2.7 concerning a single-axis stable platform system. It is desired to determine the response of this system to a random gyro drift driving function. Gyro drift is usually modeled as having a random component plus components that depend both linearly and quadratically on the acceleration of the instrument. In this problem we model only the random component. The random drift rate can be modeled in different ways, but for this problem, to simplify the analysis, let the gyro drift rate be an unbiased Gaussian white noise with

$$E\{\omega_{\text{drift}}(t)\} = 0, \qquad E\{\omega_{\text{drift}}(t)\omega_{\text{drift}}(t + \tau)\} = N\,\delta(\tau)$$

Assume that the interfering torque can be modeled as a white Gaussian noise with

$$E\{M_{\text{intf}}(t)\} = 0, \qquad E\{M_{\text{intf}}(t)M_{\text{intf}}(t + \tau)\} = M\,\delta(\tau)$$

For this problem let $F_c(p) = 1$.

Determine the expression for $E\{\omega_{\text{im}}^2(t)\}$ as a function of time, assuming M and N to be constant.

4.18 Consider a pendulum of length l with a bob of mass m as shown in Fig. 4.P4. Horizontal winds perturb the pendulum from its equilibrium position, and the perturbing force is proportional to the *relative* wind velocity on the bob, with proportionality constant a:

$$(\text{perturbing force}) = a \cdot (\text{relative wind})$$

The wind velocity is a white noise process with statistics

$$E\{w(t)\} = 0, \qquad E\{w(t)w(t + \tau)\} = b\,\delta(\tau)$$

Develop the linearized state variable stochastic differential equations for the system: for sufficiently small angles θ, write $m\ddot{x} =$ sum of forces, or equivalently, $ml^2\ddot{\theta} =$ sum of torques.

Obtain the differential equations for the elements of the appropriate covariance matrix. Determine an expression for the variance of the bob displacement, $x(t)$, when the system has reached stationary operation.
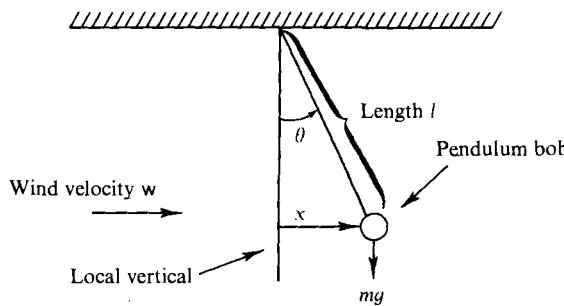
FIG. 4.P4

4.19   Given the *scalar* stochastic differential equation

$$dx(t) = [x(t)/t]\, dt + d\beta(t), \qquad t > 0$$

where

$$E\{x(t_0)\} = 0, \qquad E\{x(t_0)^2\} = a, \qquad t_0 > 0$$

and where $\beta(\cdot, \cdot)$ is Brownian motion with

$$E\{\beta(t)\} = 0, \qquad E\{[\beta(t) - \beta(t')]^2\} = q|t - t'|$$

Determine the variance of $x(t)$ for $0 < t_0 < t$.

If $y(t)$ is defined for all t as

$$y(t) = \sqrt{t}\, x(t)$$

what is $E\{y(t_2)y(t_1)\}$ for $0 < t_0 \le t_1 \le t_2$?

4.20   Consider the scalar process $x(\cdot, \cdot)$ defined on $[t_0, t_f]$ by

$$dx(t) = -(1/T)x(t)\, dt + d\beta(t)$$

where $\beta(\cdot, \cdot)$ is Brownian motion of constant diffusion parameter Q, and $x(t_0)$ is a Gaussian random variable independent of $\beta(\cdot, \cdot)$, with mean $\hat{x}_0$ and variance P_0. What must be true of $\hat{x}_0$, P_0, and Q for the $x(\cdot, \cdot)$ process to be wide-sense stationary over the interval $[t_0, t_f]$? Strict-sense stationary?

4.21   (a)   Without assuming a priori that $Q = 1$, derive the appropriate Q value for the strength of the white Gaussian noise to drive a second order shaping filter, (4-158b) or (4-158c), so as to generate a second order Markov process with autocorrelation as in (4-158a) with $\eta = 0$. Generate expressions for Q in terms of parameters σ^2, ζ, and ω_n, and show that reduction yields $Q = 1$. Obtain the result in both the time and frequency domains. Note that $\eta = 0$ iff $b^2 - a^2\omega_n{}^2 =$ (0 or $N\pi$, $N =$ integer), which specifies the location of the zero of $G(s)$ in (4-158b).

(b)   Show that η shifts the zeros of $\Psi_{xx}(\tau)$ depicted in the bottom plot of Fig. 4.22: that the first zero occurs at $\tau = (0.5\pi + \eta)/\omega_d$ and successive zeros are spaced a distance of $\Delta\tau = \pi/\omega_d$ apart, where $\omega_d = \sqrt{1 - \zeta^2}\,\omega_n$. Show that the slope $d\Psi_{xx}(\tau)/d\tau$ at $\tau = 0^+$ is given by $\sigma^2(\omega_d \tan\eta - \zeta\omega_n)$.

Thus, given an empirical autocorrelation function, the parameters required in (4-158b,c) can be established as follows. First $\sigma^2 = \Psi_{xx}(0)$. Then ω_d and η are set by the first two zeros of $\Psi_{xx}(\tau)$. The slope at $\tau = 0^+$ is used to obtain $[\zeta\omega_n] = \omega_d \tan\eta - [d\Psi_{xx}(0^+)/d\tau]/\sigma^2$. Then ζ is found by solving $\zeta/\sqrt{1 - \zeta^2} = [\zeta\omega_n]/\omega_n$. Finally, α, a, b, and c are computed as given after Eq. (4-158c).

4.22   Design the shaping filter to generate a random process having the power spectral density

$$\Psi_{xx}(\omega) = a(\omega^2 + b^2)/(\omega^4 + 4c^4)$$

from a white noise with autocorrelation $E\{w(t)w(t + \tau)\} = Q\,\delta(\tau)$. Using the state model of the filter, calculate the mean squared value of $x(t)$ for $t \in [0, \infty)$, assuming that the filter starts from rest at $t = 0$.

4.23　The shaping filter depicted in Fig. 4.P5 is meant to generate a stationary output process $x_1(\cdot, \cdot)$ in steady state operation. Show that, if $w(\cdot, \cdot)$ is zero-mean white Gaussian noise of strength Q, then in steady state,

$$E\{x_1{}^2(t)\} \to Q/[4\zeta\omega_n{}^3], \qquad E\{x_2{}^2(t)\} \to Q/[4\zeta\omega_n]$$

If this is an adequate system model, is $E\{\dot{x}_2{}^2(t)\}$ finite or not?
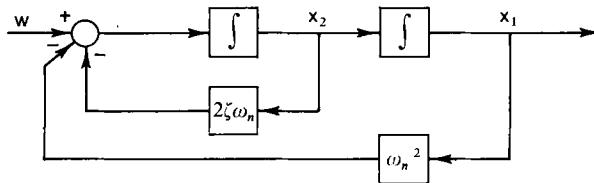
FIG. 4.P5

4.24　Design the shaping filter to generate a signal with an autocorrelation function of

$$E\{x(t)x(t + \tau)\} = K\big[(1/a)e^{-a|\tau|} - (1/b)e^{-b|\tau|}\big]$$

from an input of white noise with power spectral density value of Ψ_0.

4.25　An autocorrelation function curve-fitted to certain empirical data is of the form

$$E\{x(t)x(t + \tau)\} = \sigma^2\big[c_1 + c_2 e^{-|\tau|/T} + c_3 \cos \omega\tau\big]$$

where the positive constants c_1, c_2, c_3, σ^2, T, and ω are obtained through the curve-fitting process. Generate the state model for a shaping filter that would produce such an output process.

4.26　Two proposed models for a shaping filter to generate a process $x(\cdot, \cdot)$ with a specified autocorrelation function are as depicted in Fig. 4.P6, where $w_1(\cdot, \cdot)$ and $w_2(\cdot, \cdot)$ are white Gaussian noises of zero mean and variance kernels

$$E\{w_1(t)w_1(t + \tau)\} = q_1\,\delta(\tau), \qquad E\{w_2(t)w_2(t + \tau)\} = q_2\,\delta(\tau)$$

If these are to provide identical x process statistics in steady state operation, what is the relationship required between q_1 and q_2? If $x(\cdot, \cdot)$ is to have zero mean and mean squared value K in steady state, what values of q_1 and q_2 are required?
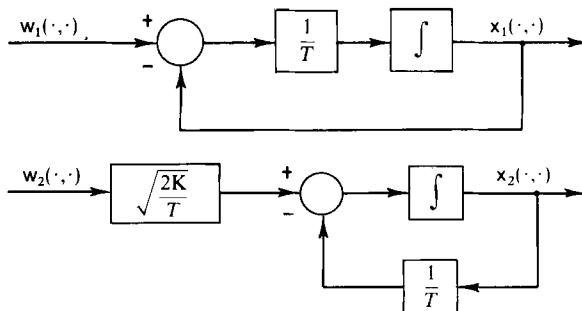
FIG. 4.P6

In terms of either q_1 or q_2, what is the autocorrelation function for the process $x(\cdot, \cdot)$ in steady state?

4.27 It can be shown that a zero-mean Gaussian process with power spectral density of $\{Q_0 + 2b\sigma^2/(\omega^2 + b^2)\}$ can be generated by summing the output of a first order lag $1/(s + b)$ driven by zero-mean white Gaussian noise of strength $(2b\sigma^2)$ with a second, independent, zero-mean white Gaussian noise of strength Q_0. Can the process also be generated as the output of a lead-lag $(s + a)/(s + b)$ driven by a single zero-mean white Gaussian noise? If so, what are the strength of the noise and the value of the parameter a?

FIG. 4.P7

4.28 Let a linear model of vertical motion of an aircraft and barometric altimeter output be as depicted in Fig. 4.P7. Vertical acceleration is integrated twice to obtain altitude, and then the altitude indicated by the barometric altimeter, $\tilde{h}$, is the output of a first order lag (because of the inherent lag in the device). Derive the state equations appropriate to this system. Now let the vertical acceleration be modeled as a wideband (approximated as white) Gaussian noise of strength q_w, with an additional time-correlated component whose autocorrelation function can be approximated as

$$E\{x(t)x(t + \tau)\} = p\exp(-\zeta\omega_n|\tau|)\cos[\omega_n(1 - \zeta^2)^{1/2}|\tau|]$$

Let the altimeter output be corrupted by

(1) a bias modeled as a process with constant samples, whose values can be described at any time t through a Gaussian random variable with mean zero and variance b^2,

(2) a wideband (approximated as white) Gaussian noise of strength R_w,

(3) an additional low frequency noise component, modeled as exponentially time-correlated noise whose mean square value is S and whose correlation time is T.

Assume all noises have zero mean.

Derive the state equations and output equation appropriate for the augmented system description.

4.29 A first order linear system is driven by white Gaussian noise. The statistics of the output $x(\cdot, \cdot)$ are, for $t \geq 0.2$ sec,

$$E\{x(t)\} = 0, \qquad E\{x^2(t)\} = 1 + e^{\cos t}(5t - 1)$$

and the statistics of the input $w(\cdot, \cdot)$ are

$$E\{w(t)\} = 0, \qquad E\{w(t)w(t + \tau)\} = [5e^{\cos t} + \sin t]\delta(\tau)$$

Find the differential equation describing the system.

CHAPTER **5**
Optimal filtering
with linear system models

5.1 INTRODUCTION

The previous chapters addressed the stochastic modeling problem: how do you develop system models that account for uncertainties in a proper, yet practical, fashion? Chapter four in particular developed practical dynamic system models in the form of linear stochastic differential or difference state equations, with associated linear output equations. Now we can exploit this background to solve a class of optimal estimation problems: equipped with such linear models and incomplete, noise-corrupted data from available sensors, how do you optimally estimate the quantities of interest to you?

Section 5.2 will formulate the problem in detail, and 5.3 will then derive the discrete-time (sampled data) optimal estimator for that problem formulation, the Kalman filter [35]. Section 5.4 investigates the characteristics of the processes within the filter structure, both to define algorithm behavior and to allow systematic event (failure) detection and adaptation. Section 5.5 delineates criteria other than that used in the derivation, with respect to which the algorithm is also optimal. Computational aspects and alternate but equivalent forms of the filter are considered in Sections 5.6 and 5.7, and stability in Section 5.8. In Sections 5.9 and 5.10 the problem formulation is extended and the resulting algorithms described. Finally, estimation for the case of continuously available measurement data and the relation of these results to Wiener filtering are explored in Sections 5.11 and 5.12.

5.2 PROBLEM FORMULATION

Assume that modeling techniques have produced an adequate system description in the form of a linear stochastic differential equation to describe the state propagation, with discrete-time noise-corrupted linear measurements available

as the system outputs. Let the *state* process $\mathbf{x}(\cdot,\cdot)$ of the system model satisfy the linear equation

$$d\mathbf{x}(t) = \mathbf{F}(t)\mathbf{x}(t)\,dt + \mathbf{B}(t)\mathbf{u}(t)\,dt + \mathbf{G}(t)\,d\boldsymbol{\beta}(t) \qquad (5\text{-}1a)$$

or, in the less rigorous white noise notation,

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \qquad (5\text{-}1b)$$

where $\mathbf{x}(\cdot,\cdot)$ is an n-vector state process, one sample of which would generate a state time history: $\mathbf{x}(t,\omega_i) = \mathbf{x}(t)$ would be the system state at time t; $\mathbf{u}(\cdot)$ is an r-vector of piecewise continuous deterministic control input functions (more general input functions are possible, but piecewise continuous is adequate for our purposes); $\mathbf{F}(\cdot)$ is an n-by-n system dynamics matrix (of piecewise continuous functions in its general form); $\mathbf{B}(\cdot)$ is an n-by-r deterministic input matrix; and $\mathbf{G}(\cdot)$ is an n-by-s noise input matrix. If (5-1a) is used, then $\boldsymbol{\beta}(\cdot,\cdot)$ is s-vector Brownian motion with statistics (for all $t, t' \in T, t \geq t'$):

$$E\{\boldsymbol{\beta}(t)\} = \mathbf{0}$$
$$E\{[\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t')][\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t')]^{\mathrm{T}}\} = \int_{t'}^{t} \mathbf{Q}(\tau)\,d\tau \qquad (5\text{-}2a)$$

with $\mathbf{Q}(\cdot)$ an s-by-s matrix of piecewise continuous functions (most generally) such that $\mathbf{Q}(t)$ is symmetric and positive semidefinite for all $t \in T$. On the other hand, if (5-1b) is used, then $\mathbf{w}(\cdot,\cdot)$ is s-vector white Gaussian noise with statistics

$$E\{\mathbf{w}(t)\} = \mathbf{0}$$
$$E\{\mathbf{w}(t)\mathbf{w}(t')^{\mathrm{T}}\} = \mathbf{Q}(t)\,\delta(t - t') \qquad (5\text{-}2b)$$

with the same description of $\mathbf{Q}(\cdot)$ as just given.

The state differential equation (5-1) is propagated forward from the *initial condition* $\mathbf{x}(t_0)$. For any particular operation of the real system, the initial state assumes a specific value $\mathbf{x}(t_0)$. However, because this value may not be known precisely a priori, it will be modeled as a random vector that is normally distributed. Thus, the description of $\mathbf{x}(t_0)$ is completely specified by the mean $\hat{\mathbf{x}}_0$ and covariance $\mathbf{P}_0$:

$$E\{\mathbf{x}(t_0)\} = \hat{\mathbf{x}}_0 \qquad (5\text{-}3a)$$
$$E\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}_0][\mathbf{x}(t_0) - \hat{\mathbf{x}}_0]^{\mathrm{T}}\} = \mathbf{P}_0 \qquad (5\text{-}3b)$$

where $\mathbf{P}_0$ is an n-by-n matrix that is symmetric and positive semidefinite. Allowing $\mathbf{P}_0$ to be positive semidefinite, instead of requiring positive definiteness, admits the case of singular $\mathbf{P}_0$: the case in which some initial states or some linear combinations of initial states are known precisely.

Measurements are available at discrete time points, $t_1, t_2, \ldots t_i, \ldots$ (often, but not necessarily, equally spaced in time), and are modeled by the relation

(for all $t_i \in T$):

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{5-4}$$

where $\mathbf{z}(\cdot,\cdot)$ is an m-vector discrete-time measurement process, one sample of which provides a particular measurement time history: $\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i$ would be the measurement numbers that become available at time t_i; $\mathbf{H}(\cdot)$ is an m-by-n measurement matrix; $\mathbf{x}(\cdot,\cdot)$ is the state vector process: $\mathbf{x}(t_i,\cdot)$ is a random vector corresponding to the state vector process at the particular time t_i; and $\mathbf{v}(\cdot,\cdot)$ is an m-vector discrete-time white Gaussian noise with statistics (for all t_i, $t_j \in T$):

$$E\{\mathbf{v}(t_i)\} = \mathbf{0} \tag{5-5a}$$

$$E\{\mathbf{v}(t_i)\mathbf{v}^{\mathrm{T}}(t_j)\} = \begin{cases} \mathbf{R}(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{5-5b}$$

In this description, $\mathbf{R}(t_i)$ is an m-by-m, symmetric, positive definite matrix for all $t_i \in T$. Positive definiteness of $\mathbf{R}(t_i)$ implies that all components of the measurement vector are corrupted by noise, and there is no linear combination of these components that would be noise-free. The measurements modeled as in (5-4) are all that we have available from the real system under consideration.

It is further assumed that $\mathbf{x}(t_0)$, $\boldsymbol{\beta}(\cdot,\cdot)$ or $\mathbf{w}(\cdot,\cdot)$, and $\mathbf{v}(\cdot,\cdot)$ are independent of each other. Since all are assumed Gaussian, this is equivalent to assuming that they are uncorrelated with each other.

It is desired to combine the measurement data taken from the actual system with the information provided by the system model and statistical description of uncertainties, in order to obtain an "optimal" estimate of the system state. In general, the "optimality" of the estimate depends upon what performance criterion is chosen. We will adopt the Bayesian point of view and seek the means of propagating the conditional probability density of the state, conditioned on the entire history of measurements taken [50]. Once this is accomplished, then the "optimal estimate" can be defined, but attention will be focused on the *entire conditional density itself*, as it embodies considerably more information in general than any single estimate value. Under the assumptions of this problem formulation, the conditional density will be shown to remain Gaussian at all times, and therefore the mean, mode, median, and essentially any logical choice of estimate based upon the conditional density will all converge upon the same estimate value.

The problem formulation just described is not the most general possible. The case of $\mathbf{R}(t_i)$ being positive semidefinite instead of positive definite, or even null in the extreme case, can be considered and will warrant subsequent attention. Furthermore, the assumed independence (uncorrelatedness) of the dynamic driving noise and the measurement noise may not provide an adequate model in some applications. For instance, in applications of optimal estimators to aircraft

navigation systems, commonly used models embody INS (inertial navigation system) noise sources in the dynamic driving noise and onboard radar uncertainties in the measurement noise: since the aircraft's own vibration affects both of those systems, there is in fact some degree of correlation among the noise sources. Admitting such correlations into the problem formulation is also possible, though again the derivation is more complicated. The given problem statement will serve as the basis for the derivation in the next section. Subsequently, extensions to the problem formulation and resulting solutions will be considered.

Before the estimator is derived, it will be convenient to introduce a new notational representation. Define a composite vector which comprises the entire measurement history to the current time, and denote it as $\mathbf{Z}(t_i)$, where

$$\mathbf{Z}(t_i) = \begin{bmatrix} \mathbf{z}(t_1) \\ \mathbf{z}(t_2) \\ \vdots \\ \mathbf{z}(t_i) \end{bmatrix} \qquad (5\text{-}6)$$

This is a vector of growing dimension: at time t_i, it is a vector random variable of dimension $(i \cdot m)$ that models the information of the entire measurement history. Its realized value, analogous to $\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i$, is $\mathbf{Z}_i$, where

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_i \end{bmatrix} \qquad (5\text{-}7)$$

This is then the history of actual measurement values obtained in a single experiment (trial). Finally, the dummy variable associated with $\mathbf{Z}(t_i)$, corresponding to ζ_i being associated with $\mathbf{z}(t_i)$, will be denoted as $\mathscr{Z}_i$.

5.3 THE DISCRETE-TIME (SAMPLED DATA) OPTIMAL ESTIMATOR: THE KALMAN FILTER

We are going to consider two measurement times, t_{i-1} and t_i, and will propagate optimal estimates from the point just after the measurement at time t_{i-1} has been incorporated into the estimate, to the point just after the measurement at time t_i is incorporated. This is depicted in Fig. 5.1 as propagating from time t_{i-1}^+ to time t_i^+.

Suppose we are at time t_{i-1} and have just taken and processed the measurement $\mathbf{z}(t_{i-1}, \omega_j) = \mathbf{z}_{i-1}$. From a Bayesian point of view, we are really interested in the probability density for $\mathbf{x}(t_{i-1})$ conditioned on the entire measurement history to that time, $f_{\mathbf{x}(t_{i-1})|\mathbf{Z}(t_{i-1})}(\xi \mid \mathbf{Z}_{i-1})$, and how this density can be propa-

FIG. 5.1 Estimate propagation. (a) Measurement $\mathbf{z}(t_{i-1}, \omega_j) = \mathbf{z}_{i-1}$ becomes available. (b) Measurement $\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i$ becomes available.

gated forward through the next measurement time to generate $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\xi|\mathbf{Z}_i)$. Once the densities are described explicitly, the optimal estimate of the state at time t_i can be determined.

To start the derivation, we will *assume* that $f_{\mathbf{x}(t_{i-1})|\mathbf{Z}(t_{i-1})}(\xi|\mathbf{Z}_{i-1})$ is a *Gaussian* conditional density:

$$
\begin{aligned}
f_{\mathbf{x}(t_{i-1})|\mathbf{Z}(t_{i-1})}(\xi|\mathbf{Z}_{i-1}) &= [(2\pi)^{n/2}|\mathbf{P}(t_{i-1}^+)|^{1/2}]^{-1}\exp\{\cdot\} \\
\{\cdot\} &= \{-\tfrac{1}{2}[\xi - \hat{\mathbf{x}}(t_{i-1}^+)]^{\mathrm{T}}\mathbf{P}^{-1}(t_{i-1}^+)[\xi - \hat{\mathbf{x}}(t_{i-1}^+)]\}
\end{aligned}
\tag{5-8}
$$

where we *define* $\hat{\mathbf{x}}(t_{i-1}^+)$ and $\mathbf{P}(t_{i-1}^+)$ to be the conditional mean and conditional covariance, respectively:

$$
\hat{\mathbf{x}}(t_{i-1}^+) \triangleq E\{\mathbf{x}(t_{i-1})|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\}
\tag{5-9}
$$

$$
\mathbf{P}(t_{i-1}^+) \triangleq E\{[\mathbf{x}(t_{i-1}) - \hat{\mathbf{x}}(t_{i-1}^+)][\mathbf{x}(t_{i-1}) - \hat{\mathbf{x}}(t_{i-1}^+)]^{\mathrm{T}}|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\}
\tag{5-10}
$$

We will be able to verify this assumption, and in fact this can be visualized as an inductive proof type of derivation, since $f_{\mathbf{x}(t_0)|\mathbf{Z}(t_0)}(\xi)$ is actually $f_{\mathbf{x}(t_0)}(\xi)$ [because the first measurement is at time t_1, so $\mathbf{Z}(t_0)$ is no measurement information at all], and $f_{\mathbf{x}(t_0)}(\xi)$ is assumed to be a Gaussian density. The following derivation considers the case from time t_{i-1} to time t_i, and combining this with the (essentially duplicate) results from t_0 to t_1 would complete an inductive proof.

Furthermore, in the process of deriving the estimator algorithm, we will be able to verify that the conditional covariance defined in (5-10) equals the unconditional covariance. In other words, the covariance recursion is *not* dependent upon the actual values of the measurements taken, and thus can be computed *without* knowledge of the realized measurement values $\mathbf{Z}_i$. For this reason, we will be able to *precompute* the time history of the covariance of the errors committed by using $\hat{\mathbf{x}}(t_i^+)$ as the optimal estimate of the state at time t_i [recall the discussion in Chapter 3: $\hat{\mathbf{x}}(t_i^+, \cdot)$ would be defined as $E\{\mathbf{x}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}(t_i, \cdot)\}$]. This will be of considerable practical significance, and will be exploited in both this chapter and the next.

Recall Fig. 5.1: we want to propagate the conditional density and associated estimate from time t_{i-1}^+, just after incorporating the measurement $\mathbf{z}(t_{i-1}, \omega_j) = \mathbf{z}_{i-1}$, to time t_i^+, just after incorporating $\mathbf{z}_i$. Let us decompose this into two

steps: (1) a time propagation from t_{i-1}^+ to t_i^-, at time t_i just before the measurement z_i is incorporated, and (2) a measurement update from t_i^- to t_i^+. To make this derivation algebraically simpler, we will at first neglect the deterministic control inputs in (5-1). Later these will be incorporated by modifying only the mean equations of the algorithm: under our assumptions, these known inputs have no effect on the spread of density functions, only on their location.

First consider the *time propagation* from t_{i-1}^+ to t_i^-. From the Bayesian point of view, we want to establish the conditional density of the state at time t_i, conditioned on the measurement history up through the previous sample time t_{i-1}: $f_{x(t_i)|Z(t_{i-1})}(\xi|Z_{i-1})$. Conceptually, we will first prove that this density is Gaussian under the assumptions of Section 5.2, and then it will be characterized completely by explicitly evaluating its mean and covariance.

By our model, $x(t_i)$ can be written as

$$\mathbf{x}(t_i) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \mathbf{w_d}(t_{i-1}) \tag{5-11}$$

where, in the context of equivalent discrete-time models,

$$\mathbf{w_d}(t_{i-1}) = \int_{t_{i-1}}^{t_i} \mathbf{\Phi}(t_i, \tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau) \tag{5-12}$$

Since (5-11) expresses $\mathbf{x}(t_i)$ as a linear combination of $\mathbf{x}(t_{i-1})$ and $\mathbf{w_d}(t_{i-1})$, $f_{x(t_i)|Z(t_{i-1})}(\xi|\mathscr{Z}_{i-1})$ will be Gaussian if we can show that $f_{x(t_{i-1}),\,w_d(t_{i-1})|Z(t_{i-1})}(\xi, \rho|\mathscr{Z}_{i-1})$ is Gaussian. By Bayes' rule,

$$f_{x(t_{i-1}),\,w_d(t_{i-1})|Z(t_{i-1})}(\xi, \rho|\mathscr{Z}_{i-1}) = \frac{f_{x(t_{i-1}),\,w_d(t_{i-1}),\,Z(t_{i-1})}(\xi, \rho, \mathscr{Z}_{i-1})}{f_{Z(t_{i-1})}(\mathscr{Z}_{i-1})}$$

Because $\mathbf{w_d}(t_{i-1})$ is independent of $\mathbf{x}(t_{i-1})$ and $\mathbf{Z}(t_{i-1})$, the numerator in this expression can be decomposed and the result recombined by another application of Bayes' rule to obtain

$$f_{x(t_{i-1}),\,w_d(t_{i-1})|Z(t_{i-1})}(\xi, \rho|\mathscr{Z}_{i-1}) = \frac{f_{x(t_{i-1}),\,Z(t_{i-1})}(\xi, \mathscr{Z}_{i-1})f_{w_d(t_{i-1})}(\rho)}{f_{Z(t_{i-1})}(\mathscr{Z}_{i-1})}$$

$$= f_{x(t_{i-1})|Z(t_{i-1})}(\xi|\mathscr{Z}_{i-1})f_{w_d(t_{i-1})}(\rho)$$

The density $f_{x(t_{i-1})|Z(t_{i-1})}(\xi|\mathscr{Z}_{i-1})$ has been assumed to be Gaussian in this induction, and $f_{w_d(t_{i-1})}(\rho)$ is Gaussian according to our dynamics model, so their product is also Gaussian. Thus, conditioned on $\mathbf{Z}(t_{i-1})$, $\mathbf{x}(t_{i-1})$ and $\mathbf{w_d}(t_{i-1})$ are jointly Gaussian, and so $f_{x(t_i)|Z(t_{i-1})}(\xi|\mathscr{Z}_{i-1})$ is in fact a Gaussian conditional density.

To specify the density completely, its mean and covariance will now be computed. The conditional mean is found by invoking the linearity of the con-

ditional expectation operator and the nonrandomness of $\boldsymbol{\Phi}(t_i, t_{i-1})$ to write

$$
\begin{aligned}
E\{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} &= E\{\boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \mathbf{w}_d(t_{i-1})|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \\
&= \boldsymbol{\Phi}(t_i, t_{i-1})E\{\mathbf{x}(t_{i-1})|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \\
&\quad + E\{\mathbf{w}_d(t_{i-1})|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\}
\end{aligned}
$$

But $\mathbf{w}_d(t_{i-1})$ is independent of $\mathbf{Z}(t_{i-1})$, so its conditional mean equals its unconditional mean, which was assumed to be zero, so

$$
E\{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} = \boldsymbol{\Phi}(t_i, t_{i-1})E\{\mathbf{x}(t_{i-1})|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \quad (5\text{-}13)
$$

Now let $\hat{\mathbf{x}}(t_i^-)$ denote the conditional mean of $\mathbf{x}(t_i)$ before the measurement $\mathbf{z}(t_i) = \mathbf{z}_i$ is taken and processed, i.e.,

$$
\hat{\mathbf{x}}(t_i^-) \triangleq E\{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \quad (5\text{-}14)
$$

In terms of this notation and that of (5-9), the conditional mean time propagation relation can be written as

$$
\hat{\mathbf{x}}(t_i^-) = \boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+) \quad (5\text{-}15)
$$

Similarly, if we define $\mathbf{P}(t_i^-)$ to be the conditional covariance of $\mathbf{x}(t_i)$ before the measurement $\mathbf{z}(t_i) = \mathbf{z}_i$ is taken and processed,

$$
\mathbf{P}(t_i^-) \triangleq E\{[\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^-)][\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^-)]^T|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \quad (5\text{-}16)
$$

then the conditional covariance propagates in time as

$$
\begin{aligned}
\mathbf{P}(t_i^-) &= \boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\boldsymbol{\Phi}^T(t_i, t_{i-1}) \\
&\quad + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\boldsymbol{\Phi}^T(t_i, \tau)\, d\tau \quad (5\text{-}17)
\end{aligned}
$$

If $\hat{\mathbf{x}}(t_i^-)$ is used as the estimate of $\mathbf{x}(t_i)$ before $\mathbf{z}_i$ is processed, then $[\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^-)]$ is the error committed by the estimator $\hat{\mathbf{x}}(t_i^-)$ for the particular measurement history realization $\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}$. Consequently $\mathbf{P}(t_i^-)$ is the conditional covariance not only of the state, but also of the error committed by using the conditional mean as an estimator of the state (this error will be shown to be zero-mean). The density function we have been seeking can now be written explicitly as

$$
\begin{aligned}
f_{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\xi}|\mathbf{Z}_{i-1}) &= [(2\pi)^{n/2}|\mathbf{P}(t_i^-)|^{1/2}]^{-1} \exp\{\cdot\} \\
\{\cdot\} &= \{-\tfrac{1}{2}[\boldsymbol{\xi} - \hat{\mathbf{x}}(t_i^-)]^T\mathbf{P}^{-1}(t_i^-)[\boldsymbol{\xi} - \hat{\mathbf{x}}(t_i^-)]\}
\end{aligned} \quad (5\text{-}18)
$$

where $\hat{\mathbf{x}}(t_i^-)$ and $\mathbf{P}(t_i^-)$ are given by (5-15) and (5-17), respectively.

Now we want to consider incorporating the measurement that becomes available at time t_i, $\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i$, so as to generate the density $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$. Repeated application of Bayes' rule will allow us to write this density in terms of three other densities, each of which can be evaluated rather easily. It is such

a desirable result that motivates and guides the particular usage of Bayes' rule that follows. For convenience and compactness, the arguments of the density functions will be omitted. Bayes' rule and the definition of $\mathbf{Z}(t_i)$ as the composite of $\mathbf{Z}(t_{i-1})$ and $\mathbf{z}(t_i)$ yield

$$
\begin{aligned}
f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)} &= \frac{f_{\mathbf{x}(t_i),\,\mathbf{Z}(t_i)}}{f_{\mathbf{Z}(t_i)}} \\[2mm]
&= \frac{f_{\mathbf{x}(t_i),\,\mathbf{z}(t_i),\,\mathbf{Z}(t_{i-1})}}{f_{\mathbf{z}(t_i),\,\mathbf{Z}(t_{i-1})}} \\[2mm]
&= \frac{f_{\mathbf{z}(t_i)|\mathbf{x}(t_i),\,\mathbf{Z}(t_{i-1})}f_{\mathbf{x}(t_i),\,\mathbf{Z}(t_{i-1})}}{f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}f_{\mathbf{Z}(t_{i-1})}} \\[2mm]
&= \frac{f_{\mathbf{z}(t_i)|\mathbf{x}(t_i),\,\mathbf{Z}(t_{i-1})}f_{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1})}f_{\mathbf{Z}(t_{i-1})}}{f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}f_{\mathbf{Z}(t_{i-1})}}
\end{aligned}
$$

Canceling like terms yields the final result,

$$
f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)} = \frac{f_{\mathbf{z}(t_i)|\mathbf{x}(t_i),\,\mathbf{Z}(t_{i-1})}f_{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1})}}{f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}} \tag{5-19}
$$

The task at hand is to evaluate each density on the right hand side of (5-19), seeking eventually to show that $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\xi|\mathscr{Z}_i)$ is Gaussian and to display it explicitly.

The second numerator term has already been established, and is given by (5-18), so let us consider the other numerator term,

$$
f_{\mathbf{z}(t_i)|\mathbf{x}(t_i),\,\mathbf{Z}(t_{i-1})}(\zeta_i|\xi,\mathscr{Z}_{i-1})
$$

According to our system model, the measurement $\mathbf{z}(t_i)$ is given by

$$
\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{5-20}
$$

We desire the density function for the random variable $\mathbf{z}(t_i)$, conditioned not only upon knowledge of the previous measurement history but also upon the fact that we *know* $\mathbf{x}(t_i)$ has assumed the realization ξ. That fact fixes the random variable $\mathbf{H}(t_i)\mathbf{x}(t_i)$ at the single known value of $\mathbf{H}(t_i)\xi$, with no uncertainty. Moreover, $\mathbf{v}(t_i)$ is independent of $\mathbf{x}(t_i)$ and $\mathbf{Z}(t_{i-1})$, and is assumed Gaussian with mean zero and covariance matrix $\mathbf{R}(t_i)$. Conditioned on $\mathbf{x}(t_i) = \xi$ and $\mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}$, $\mathbf{z}(t_i)$ is a linear combination of a known vector and a Gaussian random vector, so $f_{\mathbf{z}(t_i)|\mathbf{x}(t_i),\,\mathbf{Z}(t_{i-1})}(\zeta_i|\xi,\mathscr{Z}_{i-1})$ is a Gaussian density, completely specified by its mean and covariance. The mean is given by

$$
\begin{aligned}
E\{\mathbf{z}(t_i)|\mathbf{x}(t_i) = \xi, \mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\} &= \mathbf{H}(t_i)E\{\mathbf{x}(t_i)|\mathbf{x}(t_i) = \xi, \mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\} \\
&\quad + E\{\mathbf{v}(t_i)|\mathbf{x}(t_i) = \xi, \mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\} \\
&= \mathbf{H}(t_i)\xi \tag{5-21}
\end{aligned}
$$

The covariance matrix is:

$$E\{[\mathbf{z}(t_i) - \mathbf{H}(t_i)\boldsymbol{\xi}][\mathbf{z}(t_i) - \mathbf{H}(t_i)\boldsymbol{\xi}]^T \,|\, \mathbf{x}(t_i) = \boldsymbol{\xi}, \mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\} = \mathbf{R}(t_i) \quad (5\text{-}22)$$

Thus, we can write

$$f_{\mathbf{z}(t_i)|\mathbf{x}(t_i), \mathbf{Z}(t_{i-1})}(\boldsymbol{\zeta}_i \,|\, \boldsymbol{\xi}, \mathscr{Z}_{i-1}) = [(2\pi)^{m/2} |\mathbf{R}(t_i)|^{1/2}]^{-1} \exp\{\cdot\}$$
$$\{\cdot\} = \{-\tfrac{1}{2}[\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\boldsymbol{\xi}]^T \mathbf{R}^{-1}(t_i)[\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\boldsymbol{\xi}]\} \quad (5\text{-}23)$$

Having evaluated the numerator terms in (5-19), we now consider the denominator, $f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\zeta}_i|\mathscr{Z}_{i-1})$. The measurement $\mathbf{z}(t_i)$ is again described as in (5-20), but now we are conditioning only on knowledge of the previous time history of measurements. First we want to show that $f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\zeta}_i|\mathscr{Z}_{i-1})$ is Gaussian. Since $\mathbf{z}(t_i)$ is a linear combination of $\mathbf{x}(t_i)$ and $\mathbf{v}(t_i)$, we will be able to achieve this objective if we can show that, conditioned on $\mathbf{Z}(t_{i-1})$, $\mathbf{x}(t_i)$ and $\mathbf{v}(t_i)$ are jointly Gaussian. Bayes' rule yields:

$$f_{\mathbf{x}(t_i), \mathbf{v}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\xi}, \boldsymbol{\eta} \,|\, \mathscr{Z}_{i-1}) = f_{\mathbf{v}(t_i)|\mathbf{x}(t_i), \mathbf{Z}(t_{i-1})}(\boldsymbol{\eta} \,|\, \boldsymbol{\xi}, \mathscr{Z}_{i-1})$$
$$\cdot f_{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\xi} \,|\, \mathscr{Z}_{i-1})$$

But $\mathbf{v}(t_i)$ is independent of $\mathbf{x}(t_i)$ and $\mathbf{Z}(t_{i-1})$, so this becomes

$$f_{\mathbf{x}(t_i), \mathbf{v}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\xi}, \boldsymbol{\eta} \,|\, \mathscr{Z}_{i-1}) = f_{\mathbf{v}(t_i)}(\boldsymbol{\eta}) f_{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\xi} \,|\, \mathscr{Z}_{i-1})$$

The two separate densities on the right hand side of this expression are each Gaussian, so their product is Gaussian, and thus $f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\zeta}_i|\mathscr{Z}_{i-1})$ is itself Gaussian. The mean is calculated as:

$$E\{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\} = \mathbf{H}(t_i)E\{\mathbf{x}(t_i)|\mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\}$$
$$+ E\{\mathbf{v}(t_i)|\mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\}$$
$$= \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) \quad (5\text{-}24)$$

The covariance is computed as

$$E\{[\mathbf{z}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)][\mathbf{z}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]^T | \mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\}$$
$$= E\{[\mathbf{H}(t_i)\mathbf{x}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) + \mathbf{v}(t_i)]$$
$$\cdot [\mathbf{H}(t_i)\mathbf{x}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) + \mathbf{v}(t_i)]^T | \mathbf{Z}(t_{i-1}) = \mathscr{Z}_{i-1}\}$$
$$= \mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i) \quad (5\text{-}25)$$

Since $f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\zeta}_i|\mathscr{Z}_{i-1})$ is a Gaussian density, we can now write:

$$f_{\mathbf{z}(t_i)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\zeta}_i|\mathscr{Z}_{i-1}) = [(2\pi)^{m/2} |\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)|^{1/2}]^{-1} \exp\{\cdot\}$$
$$\{\cdot\} = \{-\tfrac{1}{2}[\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]^T [\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)]^{-1}[\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]\}$$
$$(5\text{-}26)$$

At this point, we have written (5-19) to generate a measurement update expression for evaluating $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathscr{Z}_i)$, and we have depicted each of the
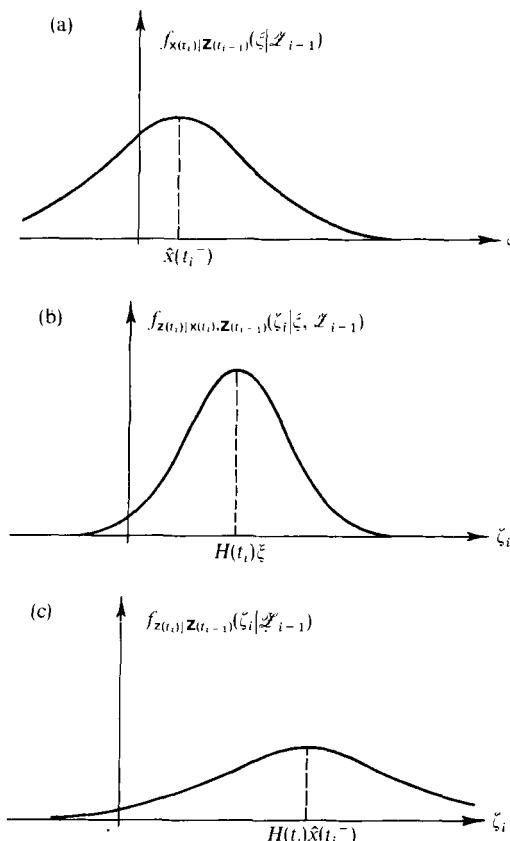
FIG. 5.2   Evaluation of densities for use in Bayes' rule. (a) Variance $= P(t_i^-)$. (b) Variance $= R(t_i)$. (c) Variance $= H(t_i)P(t_i^-)H(t_i) + R(t_i)$.

separate Gaussian densities explicitly. This is portrayed graphically in Fig. 5.2. Substituting (5-18), (5-23), and (5-26) into (5-19) yields

$$f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}\,|\,\mathscr{Z}_i) = \frac{|\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i)|^{1/2}}{(2\pi)^{n/2}|\mathbf{P}(t_i^-)|^{1/2}|\mathbf{R}(t_i)|^{1/2}}\,\exp\{-\tfrac{1}{2}(\cdot)\}$$

$$\begin{aligned}
(\cdot) = {} & [\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\boldsymbol{\xi}]^{\mathrm{T}}\mathbf{R}(t_i)^{-1}[\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\boldsymbol{\xi}] \\
& + [\boldsymbol{\xi} - \hat{\mathbf{x}}(t_i^-)]^{\mathrm{T}}\mathbf{P}(t_i^-)^{-1}[\boldsymbol{\xi} - \hat{\mathbf{x}}(t_i^-)] \\
& - [\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]^{\mathrm{T}}[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i)]^{-1}[\boldsymbol{\zeta}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]
\end{aligned} \tag{5-27}$$

It is not immediately evident that (5-27) is in fact of Gaussian form: the three separate determinant terms would have to be equivalent to a single determinant square root in the denominator of the leading coefficient, and the sum of the

three quadratics in the exponential would have to be equivalent to a single quadratic form. We will demonstrate the quadratic form equivalency, and the manipulation of determinants is left to Problem 5.4 at the end of the chapter.

To achieve the desired result, we will require use of "the matrix inversion lemma," valid for positive definite $\mathbf{P}$ and $\mathbf{R}$:

$$[\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1} = \mathbf{P} - \mathbf{P}\mathbf{H}^T[\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{H}\mathbf{P} \qquad (5\text{-}28)$$

This lemma is important enough to warrant proving; one such proof is outlined in Problem 5.2. It is of special interest to us because the left hand side involves inversion of n-by-n matrices, whereas the right hand side requires m-by-m matrix inversion: in most problems of interest, m will be significantly less than n. A more general lemma, admitting positive semidefinite $\mathbf{P}$, can be proven to yield

$$[\mathbf{I} + \mathbf{P}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{P} = \mathbf{P} - \mathbf{P}\mathbf{H}^T[\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{H}\mathbf{P} \qquad (5\text{-}28')$$

but we will not need this generality here. In fact, in the ensuing proof, we will *assume* $\mathbf{P}(t_i{}^-)$ to be positive definite and at the end of the development we will return to establish under what conditions the assumption is valid. Once (5-28) is proven, it is straightforward to generate two other useful matrix identities (see Problem 5.3) as well:

$$[\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{P}\mathbf{H}^T[\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}]^{-1} \qquad (5\text{-}29)$$

$$\mathbf{H}[\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{H}^T = \mathbf{R} - \mathbf{R}[\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{R} \qquad (5\text{-}30)$$

What follows is somewhat difficult to motivate, in much the same way as scalar completion of squares is, except by keeping the overall objective firmly in mind: to generate a single quadratic form from the sum of three quadratics in (5-27). Through expanding, exploiting algebraic identities, and regrouping, we seek to combine all terms into a single quadratic. To make the algebra more tractable, we will omit the time notation, and denote $\hat{\mathbf{x}}(t_i{}^-)$ and $\mathbf{P}(t_i{}^-)$ by $\hat{\mathbf{x}}^-$ and $\mathbf{P}^-$, respectively.

REDUCTION TO A SINGLE QUADRATIC FORM   First expand the terms denoted by $(\cdot)$ in (5-27) to get the sum of 12 terms, which can be combined conveniently as (recalling that the transpose of a scalar is just the scalar itself)

$$\begin{aligned}
(\cdot) = {}& \boldsymbol{\xi}^T[\mathbf{P}^{-\,-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]\boldsymbol{\xi} - 2\boldsymbol{\xi}^T[\mathbf{P}^{-\,-1}\hat{\mathbf{x}}^- + \mathbf{H}^T\mathbf{R}^{-1}\boldsymbol{\zeta}_i] \\
& + \boldsymbol{\zeta}_i{}^T[\mathbf{R}^{-1} - (\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}]\boldsymbol{\zeta}_i + 2\hat{\mathbf{x}}^{-T}\mathbf{H}^T[\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]^{-1}\boldsymbol{\zeta}_i \\
& + \hat{\mathbf{x}}^{-T}[\mathbf{P}^{-\,-1} - \mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}]\hat{\mathbf{x}}^-
\end{aligned}$$

Consider the third term. The matrix which appears in it can be obtained by premultiplying and postmultiplying (5-30) by $\mathbf{R}^{-1}$ (which exists since $\mathbf{R}$ is positive definite):

$$\begin{aligned}
\mathbf{R}^{-1}\mathbf{H}[\mathbf{P}^{-\,-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{R}^{-1} &= \mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{R}[\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{R}\mathbf{R}^{-1} \\
&= \mathbf{R}^{-1} - [\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]^{-1}
\end{aligned}$$

so that the third term can be rewritten as

$$\{ + \boldsymbol{\zeta}_i^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H} [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{\zeta}_i \}$$

Now the objective is to operate on the fourth and fifth terms so as to express them in terms of a quadratic form as $\mathbf{x}_1^{\mathrm{T}} [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]^{-1} \mathbf{x}_2$, with $\mathbf{x}_1$ and $\mathbf{x}_2$ some n-vectors, so that subsequent combination of terms will be possible.

Now consider the fourth term. If (5-29) is premultiplied by $\mathbf{P}^{-\,-1}$ (which exists since $\mathbf{P}^-$ is assumed positive definite), we get

$$\mathbf{P}^{-\,-1} [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} = \mathbf{P}^{-\,-1} \mathbf{P}^- \mathbf{H}^{\mathrm{T}} (\mathbf{H} \mathbf{P}^- \mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}$$
$$= \mathbf{H}^{\mathrm{T}} [\mathbf{H} \mathbf{P}^- \mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}$$

Thus, the fourth term can be rewritten as

$$\{ + 2\hat{\mathbf{x}}^{-\mathrm{T}} \mathbf{P}^{-\,-1} [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{\zeta}_i \}$$

Finally, look at the fifth term. If (5-28) is premultiplied and postmultiplied by $\mathbf{P}^{-\,-1}$, we get

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

Thus the fifth term can be put into the form

$$\{ + \hat{\mathbf{x}}^{-\mathrm{T}} \mathbf{P}^{-\,-1} [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]^{-1} \mathbf{P}^{-\,-1} \hat{\mathbf{x}}^- \}$$

Now the third, fourth, and fifth terms in the original expansion can be combined so as to write the expansion equivalently as

$$(\cdot) = \boldsymbol{\xi}^{\mathrm{T}} [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}] \boldsymbol{\xi} - 2 \boldsymbol{\xi}^{\mathrm{T}} [\mathbf{P}^{-\,-1} \hat{\mathbf{x}}^- + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{\zeta}_i]$$
$$+ [\boldsymbol{\zeta}_i^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H} + \hat{\mathbf{x}}^{-\mathrm{T}} \mathbf{P}^{-1}] [\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]^{-1} [\mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{\zeta}_i + \mathbf{P}^{-1} \hat{\mathbf{x}}^-]$$

To simplify the remaining algebra, define the n-vector $\mathbf{a}$ and the n-by-n matrix $\mathbf{A}$ as

$$\mathbf{a} \triangleq [\mathbf{P}^{-\,-1} \hat{\mathbf{x}}^- + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{\zeta}_i] \qquad \mathbf{A} \triangleq [\mathbf{P}^{-\,-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}]$$

In terms of this notation, the expansion becomes

$$(\cdot) = \boldsymbol{\xi}^{\mathrm{T}} \mathbf{A} \boldsymbol{\xi} - 2 \boldsymbol{\xi}^{\mathrm{T}} \mathbf{a} + \mathbf{a}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{a}$$

Again motivated by the desired to achieve a single quadratic form, we can write this equivalently as

$$(\cdot) = \boldsymbol{\xi}^{\mathrm{T}} \mathbf{A} \boldsymbol{\xi} - 2 \boldsymbol{\xi}^{\mathrm{T}} \mathbf{A} \mathbf{A}^{-1} \mathbf{a} + \mathbf{a}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{A} \mathbf{A}^{-1} \mathbf{a}$$
$$= (\boldsymbol{\xi} - \mathbf{A}^{-1} \mathbf{a})^{\mathrm{T}} \mathbf{A} (\boldsymbol{\xi} - \mathbf{A}^{-1} \mathbf{a})$$

This is the single quadratic form we have been seeking.   ∎

Combined with a similar development for the determinant terms, the preceding reduction has shown that the conditional probability density $f_{\mathbf{x}(t_i) | \mathbf{Z}(t_i)}(\boldsymbol{\xi} | \mathbf{Z}_i)$ is indeed a Gaussian density, with mean $(\mathbf{A}^{-1} \mathbf{a})$ and covariance $\mathbf{A}^{-1}$. Consistent with the previous definition of $\hat{\mathbf{x}}(t_{i-1}^+)$, the mean of this conditional density is denoted as $\hat{\mathbf{x}}(t_i^+)$:

$$\hat{\mathbf{x}}(t_i^+) \triangleq E\{\mathbf{x}(t_i) | \mathbf{Z}(t_i) = \mathbf{Z}_i\} = \mathbf{A}^{-1} \mathbf{a}$$
$$= [\mathbf{P}(t_i^-)^{-1} + \mathbf{H}^{\mathrm{T}}(t_i) \mathbf{R}^{-1}(t_i) \mathbf{H}(t_i)]^{-1} [\mathbf{P}(t_i^-)^{-1} \hat{\mathbf{x}}(t_i^-) + \mathbf{H}^{\mathrm{T}}(t_i) \mathbf{R}^{-1}(t_i) \mathbf{z}_i]$$

$$(5\text{-}31)$$

Similarly, the covariance is denoted as $\mathbf{P}(t_i{}^+)$:

$$
\begin{aligned}
\mathbf{P}(t_i{}^+) &\triangleq E\{[\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i{}^+)][\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i{}^+)]^{\mathrm{T}} | \mathbf{Z}(t_i) = \mathbf{Z}_i\} = \mathbf{A}^{-1} \\
&= [\mathbf{P}(t_i{}^-)^{-1} + \mathbf{H}^{\mathrm{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)]^{-1}
\end{aligned}
\tag{5-32}
$$

In terms of these statistics, the density $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$ can be written explicitly as:

$$
f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i) = [(2\pi)^{n/2}|\mathbf{P}(t_i{}^+)|^{1/2}]^{-1}\exp\{\cdot\}
$$

$$
\{\cdot\} = \{-\tfrac{1}{2}[\boldsymbol{\xi} - \hat{\mathbf{x}}(t_i{}^+)]^{\mathrm{T}}\mathbf{P}(t_i{}^+)^{-1}[\boldsymbol{\xi} - \hat{\mathbf{x}}(t_i{}^+)]\}
\tag{5-33}
$$

This is portrayed in Fig. 5.3. As indicated by the notation, the conditional mean $\hat{\mathbf{x}}(t_i{}^+)$ is chosen as the optimal estimate. Not only is it the conditional mean, but also the conditional mode: it maximizes the conditional density of $\mathbf{x}(t_i)$ conditioned on the entire measurement history (i.e., it is more probable to be in the interval between $[\hat{\mathbf{x}}(t_i{}^+) - \boldsymbol{\varepsilon}]$ and $[\hat{\mathbf{x}}(t_i{}^+) + \boldsymbol{\varepsilon}]$ than any equivalent-sized region of $\boldsymbol{\xi}$). It is also the conditional median and satisfies essentially any criterion of optimality once $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$ has been established. As mentioned previously, if we do in fact use $\hat{\mathbf{x}}(t_i{}^+)$ as the optimal estimate, then $\mathbf{P}(t_i{}^+)$ is not only the state covariance but also the covariance of the error committed by that estimate of the state value.
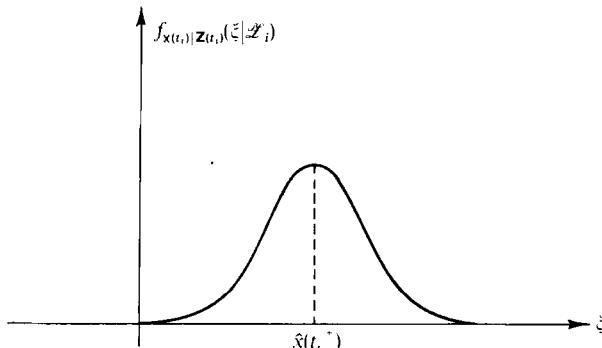
FIG. 5.3   Conditional probability density after measurement incorporation. Variance = $P(t_i{}^+)$.

Although (5-31) and (5-32) are valid expressions, they involve inversions of n-by-n matrices, where n is the dimension of the state vector. If m, the dimension of the measurement vector, is significantly smaller than n (as is often the case), then the matrix inversion lemma can yield equivalent but more efficient expressions that require only m-by-m inversions. Substituting (5-28) and (5-29) into (5-31) yields:

$$
\hat{\mathbf{x}}(t_i{}^+) = [\mathbf{P}^- - \mathbf{P}^-\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}]\mathbf{P}^{- -1}\hat{\mathbf{x}}^- + [\mathbf{P}^-\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^-\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}]\mathbf{z}_i
$$

Regrouping these terms and applying (5-28) directly to (5-32) yields the measurement update equations as:

$$\hat{\mathbf{x}}(t_i{}^+) = \hat{\mathbf{x}}(t_i{}^-) + \mathbf{P}(t_i{}^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i{}^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)]^{-1}[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i{}^-)]$$
$$(5\text{-}34)$$

$$\mathbf{P}(t_i{}^+) = \mathbf{P}(t_i{}^-) - \mathbf{P}(t_i{}^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i{}^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)]^{-1}\mathbf{H}(t_i)\mathbf{P}(t_i{}^-) \qquad (5\text{-}35)$$

Recall that this derivation was based upon the assumption that $\mathbf{P}(t_i{}^-)$ was positive definite. Let us investigate the conditions under which this is valid. First, we will determine if $\mathbf{P}(t_{i-1}^+)$ being positive definite is sufficient to make $\mathbf{P}(t_i{}^+)$ positive definite. In (5-17), if $\mathbf{P}(t_{i-1}^+)$ is assumed positive definite, then $\mathbf{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\mathbf{\Phi}^T(t_i, t_{i-1})$ is also positive definite by the properties of state transition matrices; the integral term is at worst positive semidefinite, so $\mathbf{P}(t_i{}^-)$ is positive definite if $\mathbf{P}(t_{i-1}^+)$ is. In generating $\mathbf{P}(t_i{}^+)$ from $\mathbf{P}(t_i{}^-)$ as in (5-35), there would seem to be some question about preserving positive definiteness because a term is subtracted from $\mathbf{P}(t_i{}^-)$ to obtain $\mathbf{P}(t_i{}^+)$. However, if the equivalent expression (5-31) is considered, this preservation becomes evident. Since $\mathbf{P}(t_i{}^-)$ is assumed positive definite, $\mathbf{P}(t_i{}^-)^{-1}$ is also positive definite. Added to this is the term $\mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)$, which is positive semidefinite [since $\mathbf{R}(t_i)$ is assumed positive definite, $\mathbf{R}(t_i)^{-1}$ is positive definite, and so $\mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)$ is an n-by-n matrix of rank at most m], so their sum $\mathbf{P}(t_i{}^+)^{-1}$ is positive definite, and so is its inverse, $\mathbf{P}(t_i{}^+)$.

Thus, we can conclude that once $\mathbf{P}(t_i{}^-)$ or $\mathbf{P}(t_i{}^+)$ becomes positive definite, the covariances will remain positive definite from that time forward (although they may asymptotically approach singularity). For that reason, look at the initial time interval and determine under what conditions

$$\mathbf{P}(t_1{}^-) = \mathbf{\Phi}(t_1, t_0)\mathbf{P}_0\mathbf{\Phi}^T(t_1, t_0) + \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\mathbf{\Phi}^T(t_1, \tau)\, d\tau$$

is positive definite. Two sufficient (not necessary) conditions would be if $\mathbf{P}_0$ were positive definite or if the integral term were separately positive definite [i.e., $\mathbf{Q}(t)$ is positive definite for all $t \in [t_0, t_1)$ and the system description is completely controllable from the points of entry of the dynamic driving noise]. Neither of these are very restrictive assumptions, yet we really only require $\mathbf{P}(t_1{}^-)$ itself to be positive definite.

The derivation just presented was not the most general possible. For instance, $\mathbf{R}(t_i)$ need not be positive definite for the algorithm to operate properly (as long as $[\mathbf{H}(t_i)\mathbf{P}(t_i{}^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)]$ is always invertible), even though this was required in our derivation. Nevertheless, the assumption made does encompass the vast majority of applications of practical interest.

To complete the derivation, let us add the effects of deterministic control inputs. As described previously, the *only* change in the estimator algorithm is that the state estimate (conditional mean) time propagation relation (5-15)

becomes

$$\hat{\mathbf{x}}(t_i^{\,-}) = \boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^{\,+}) + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$$

Note that the discrete-time (sampled data) *Kalman filter algorithm* just derived entails the time propagation and measurement updating of conditional mean and covariance equations. However, because all probability densities of interest have been shown to be Gaussian, this algorithm does in fact portray the entire conditional density of the state conditioned on the measurements taken: the Bayesian objective of propagating *all* probability information has been fulfilled.

To summarize the algorithm, the optimal state estimate is *propagated* from measurement time t_{i-1} to measurement time t_i by the relations

$$\hat{\mathbf{x}}(t_i^{\,-}) = \boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^{\,+}) + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \tag{5-36}$$

$$\begin{aligned}\mathbf{P}(t_i^{\,-}) &= \boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^{\,+})\boldsymbol{\Phi}^{\mathsf{T}}(t_i, t_{i-1}) \\ &+ \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathsf{T}}(\tau)\boldsymbol{\Phi}^{\mathsf{T}}(t_i, \tau)\,d\tau \end{aligned}\tag{5-37}$$

At measurement time t_i, the measurement $\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i$ becomes available. The estimate is *updated* by defining the Kalman filter gain $\mathbf{K}(t_i)$ and employing it in both the mean and covariance relations:

$$\mathbf{K}(t_i) = \mathbf{P}(t_i^{\,-})\mathbf{H}^{\mathsf{T}}(t_i)\left[\mathbf{H}(t_i)\mathbf{P}(t_i^{\,-})\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i)\right]^{-1} \tag{5-38}$$

$$\hat{\mathbf{x}}(t_i^{\,+}) = \hat{\mathbf{x}}(t_i^{\,-}) + \mathbf{K}(t_i)\left[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^{\,-})\right] \tag{5-39}$$

$$\mathbf{P}(t_i^{\,+}) = \mathbf{P}(t_i^{\,-}) - \mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i^{\,-}) \tag{5-40}$$

The initial conditions for the recursion are given by

$$\hat{\mathbf{x}}(t_0) = E\{\mathbf{x}(t_0)\} = \hat{\mathbf{x}}_0 \tag{5-41}$$

$$\mathbf{P}(t_0) = E\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}_0][\mathbf{x}(t_0) - \hat{\mathbf{x}}_0]^{\mathsf{T}}\} = \mathbf{P}_0 \tag{5-42}$$

Figure 5.4 is a block diagram portrayal of the algorithm. The mathematical system model inherently in the filter structure generates $\hat{\mathbf{x}}(t_i^{\,-})$, the best prediction of the state at time t_i before the measurement at time t_i, $\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i$, is
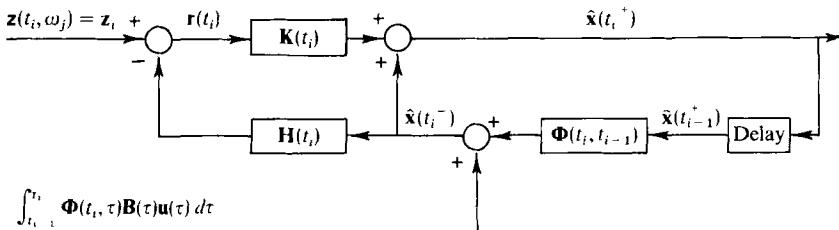
FIG. 5.4   Sampled-data Kalman filter block diagram.

processed. Moreover, this same system model allows generation of $[\mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^{-})]$, which is the best prediction of what the measurement at time t_i will be before it is actually taken [recall (5-24)]. The input to the algorithm is $\mathbf{z}_i$, the realized value of the measurement $\mathbf{z}(t_i)$. The measurement *residual* $\mathbf{r}(t_i)$ is then generated as the difference between the true measurement value $\mathbf{z}_i$ and the best prediction of it before it is actually taken:

$$\mathbf{r}(t_i) = \mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^{-}) \tag{5-43}$$

(Many term this quantity the *innovations* and reserve the name "residual" for the quantity $[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^{+})]$, which does not appear explicitly in the algorithm; we choose to retain the name residual because of the important procedure known as "residual monitoring," in which the sequence of $\mathbf{r}(t_i)$ values are monitored for adaptive purposes.) Then the residual is passed through an optimal weighting matrix $\mathbf{K}(t_i)$ to generate a correction term to be added to $\hat{\mathbf{x}}(t_i^{-})$ to obtain $\hat{\mathbf{x}}(t_i^{+})$: the algorithm has a predictor–corrector structure.

To specify a Kalman filter algorithm completely, we need to define both the *structure* of the system model and a statistical description of the *uncertainties* in the model. The structure is established by $\mathbf{F}(t)$ or $\mathbf{\Phi}(t_i, \tau)$, $\mathbf{B}(t)$, $\mathbf{G}(t)$, and $\mathbf{H}(t_i)$ for all times of interest, and the uncertainties are specified by $\hat{\mathbf{x}}_0$, $\mathbf{P}_0$, and the time histories of $\mathbf{Q}(t)$ and $\mathbf{R}(t_i)$.

EXAMPLE 5.1   Recall the example of being lost at sea in Section 1.5 of the first chapter. Just after time t_2, after the trained navigator's measurement was incorporated, the state estimate and variance were established as $\hat{x}(t_2^{+})$ and $P(t_2^{+}) = \sigma_x^2(t_2^{+})$. The dynamics model was given by (1-10) as

$$\dot{x}(t) = u + w(t)$$

with u constant and $w(\cdot, \cdot)$ described as a zero-mean white Gaussian noise of strength

$$E\{w(t)w(t + \tau)\} = \sigma_w^2\delta(\tau)$$

Thus, we can identify $\mathbf{F} = 0$ so $\mathbf{\Phi} = 1$, $\mathbf{B} = 1$, $\mathbf{G} = 1$, and $\mathbf{Q} = \sigma_w^2$.

The time propagation equation for the state estimate would be

$$\hat{x}(t_3^{-}) = \Phi(t_3, t_2)\hat{x}(t_2^{+}) + \int_{t_2}^{t_3} \Phi(t_3, \tau)B(\tau)u(\tau)\,d\tau$$

$$= 1 \cdot \hat{x}(t_2^{+}) + \int_{t_2}^{t_3} 1 \cdot 1 \cdot u\,d\tau$$

$$= \hat{x}(t_2^{+}) + u[t_3 - t_2]$$

This is the result quoted in (1-11). Similarly, the variance time propagation is as delineated in (1-12),

$$P(t_3^{-}) = \Phi(t_3, t_2)P(t_2^{+})\Phi^{\mathsf{T}}(t_3, t_2) + \int_{t_2}^{t_3} \Phi(t_3, \tau)G(\tau)Q(\tau)G^{\mathsf{T}}(\tau)\Phi^{\mathsf{T}}(t_3, \tau)\,d\tau$$

$$= 1 \cdot P(t_2^{+}) \cdot 1 + \int_{t_2}^{t_3} 1 \cdot 1 \cdot \sigma_w^2 \cdot 1 \cdot 1\,d\tau$$

$$= P(t_2^{+}) + \sigma_w^2[t_3 - t_2]$$

At time t_3, a measurement becomes available, modeled as

$$z(t_3) = x(t_3) + v(t_3)$$

where $v(\cdot, \cdot)$ is a white Gaussian discrete-time noise process of mean zero and variance $\sigma_{z_3}^2$: so $H = 1$, $R = \sigma_{z_3}^2$. Based on this model, the measurement update relations become

$$
\begin{aligned}
K(t_3) &= P(t_3{}^-)H^T(t_3)[H(t_3)P(t_3{}^-)H^T(t_3) + R(t_3)]^{-1} \\
&= P(t_3{}^-)/[P(t_3{}^-) + \sigma_{z_3}^2] \\
\hat{x}(t_3{}^+) &= \hat{x}(t_3{}^-) + K(t_3)[z_3 - H(t_3)\hat{x}(t_3{}^-)] \\
&= \hat{x}(t_3{}^-) + K(t_3)[z_3 - \hat{x}(t_3{}^-)] \\
P(t_3{}^+) &= P(t_3{}^-) - K(t_3)H(t_3)P(t_3{}^-) \\
&= P(t_3{}^-) - K(t_3)P(t_3{}^-)
\end{aligned}
$$

These are identical to (1-13)–(1-15).   ■

The Kalman filter time propagation equations given by (5-36) and (5-37) are in a convenient form if the problem under consideration is modeled by time-invariant dynamics and dynamic driving noise with stationary statistics (F, B, G, and Q all constants) with a fixed measurement sample period. In this case, $\Phi(t_i, t_{i-1})$ and $\int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)G(\tau)Q(\tau)G^T(\tau)\Phi^T(t_i, \tau)\,d\tau$ are the same for every sample period and need only be computed once. Moreover, if the deterministic inputs are held constant over each sample period (as provided by a digital controller operating at the same iteration rate), then $u(\tau) = u(t_{i-1})$ for all $\tau \in [t_{i-1}, t_i)$, and

$$
\int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)B(\tau)u(t_{i-1})\,d\tau = \left[\int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)B(\tau)\,d\tau \right] u(t_{i-1})
$$

and the bracketed term is also constant from sample to sample.

However, if the system or statistics are time varying, the time propagation equations would be more efficiently expressed in differential equation form. To generate this form, define $\hat{x}(t/t_{i-1})$ and $P(t/t_{i-1})$ for any $t \in [t_{i-1}, t_i)$ as the conditional mean and covariance conditioned on the measurements taken up to that time, i.e., up through $z(t_{i-1}, \omega_j) = z_{i-1}$:

$$
\hat{x}(t/t_{i-1}) = \Phi(t, t_{i-1})\hat{x}(t_{i-1}^+) + \int_{t_{i-1}}^t \Phi(t, \tau)B(\tau)u(\tau)\,d\tau \tag{5-44}
$$

$$
P(t/t_{i-1}) = \Phi(t, t_{i-1})P(t_{i-1}^+)\Phi^T(t, t_{i-1})
$$
$$
+ \int_{t_{i-1}}^t \Phi(t, \tau)G(\tau)Q(\tau)G^T(\tau)\Phi^T(t, \tau)\,d\tau \tag{5-45}
$$

Differentiating these yields

$$
\dot{\hat{x}}(t/t_{i-1}) = F(t)\hat{x}(t/t_{i-1}) + B(t)u(t) \tag{5-46}
$$

$$
\dot{P}(t/t_{i-1}) = F(t)P(t/t_{i-1}) + P(t/t_{i-1})F^T(t) + G(t)Q(t)G^T(t) \tag{5-47}
$$

which would then be integrated over the interval from t_{i-1} to t_i, starting from the initial conditions

$$
\hat{x}(t_{i-1}/t_{i-1}) = \hat{x}(t_{i-1}^+) \tag{5-48a}
$$

$$
P(t_{i-1}/t_{i-1}) = P(t_{i-1}^+) \tag{5-48b}
$$

Integrating these to time t_i (as by fourth order Runge–Kutta technique or by sequentially integrating over partitions of this interval by a first order method) would yield $\hat{\mathbf{x}}(t_i^-)$ and $\mathbf{P}(t_i^-)$.

In some cases the dynamics model is itself a discrete-time model, for instance an equivalent discrete-time model as discussed in the previous chapter. Let us replace the dynamics model of (5-1) and (5-2) with the linear stochastic difference equation

$$\mathbf{x}(t_i) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \mathbf{B}_d(t_{i-1})\mathbf{u}(t_{i-1}) + \mathbf{G}_d(t_{i-1})\mathbf{w}_d(t_{i-1}) \qquad (5\text{-}49)$$

where $\mathbf{w}_d(\cdot, \cdot)$ is a discrete-time zero-mean white Gaussian noise sequence with covariance kernel

$$E\{\mathbf{w}_d(t_i)\mathbf{w}_d^\mathrm{T}(t_j)\} = \begin{cases} \mathbf{Q}_d(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \qquad (5\text{-}50)$$

(Note that in the equivalent discrete model formulation, $\mathbf{G}_d$ was assumed to be the identity matrix.) The only change to the Kalman filter algorithm is that the time propagation equations (5-36) and (5-37) are replaced by

$$\hat{\mathbf{x}}(t_i^-) = \mathbf{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+) + \mathbf{B}_d(t_{i-1})\mathbf{u}(t_{i-1}) \qquad (5\text{-}51)$$

$$\mathbf{P}(t_i^-) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\mathbf{\Phi}^\mathrm{T}(t_i, t_{i-1}) + \mathbf{G}_d(t_{i-1})\mathbf{Q}_d(t_{i-1})\mathbf{G}_d^\mathrm{T}(t_{i-1}) \qquad (5\text{-}52)$$

To specify a Kalman filter of this form completely, we again must depict both the structure $[\mathbf{\Phi}(t_i, t_{i-1}), \mathbf{B}_d(t_{i-1}), \mathbf{G}_d(t_{i-1}), \mathbf{H}(t_i)$ for all times of interest$]$ and uncertainties $[\hat{\mathbf{x}}_0, \mathbf{P}_0,$ and $\mathbf{Q}_d(t_{i-1})$ and $\mathbf{R}(t_i)$ for all times$]$ of the model.

EXAMPLE 5.2   We now consider an example based on a scalar dynamics model: a simple representation of a gyro on test. (This example will be reexamined throughout Chapter 5.) Gyros are subject to long term drifts, and we would like to estimate the drift rate from laboratory data. Assume that gyro drift rate can be adequately modeled as a stationary exponentially time-correlated Gaussian process (in fact, this is a rather good model for the dominant drift effects). To keep the problem restricted to one state variable, we will further assume that we can measure instantaneous drift rate. Thus, though not totally realistic, this problem can be viewed as a portion of a more realistic problem, and we seek to exploit its simple form to illustrate the use of the estimator algorithm.

Figure 5.5 depicts the system model. Gyro drift rate is the state process $\mathbf{x}(\cdot, \cdot)$, and since it is an exponentially time-correlated Gaussian process, it is shown as the output of a first order shaping filter (first order lag) driven by white Gaussian noise $\mathbf{w}(\cdot, \cdot)$. The shaping filter break fre-
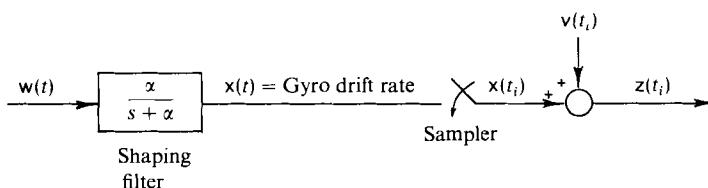
FIG. 5.5   System model for gyro on test example.

quency α is set at 1 rad/hr, i.e., the correlation time of the x process is one hour (reasonable for state-of-the-art gyros), and the statistics of $w(\cdot, \cdot)$ are

$$E\{w(t)\} = 0$$
$$E\{w(t)w(t + \tau)\} = Q\,\delta(\tau), \qquad Q = 2 \quad \deg^2/hr$$

The units of Q seem strange at first since w is in units of deg/hr, but are valid because $\delta(\tau)$ carries units of $(time)^{-1}$. It is assumed that sampled data measurements are taken every 0.25 hr, modeled as

$$z(t_i) = x(t_i) + v(t_i)$$

where $v(\cdot, \cdot)$ is a discrete-time zero-mean white Gaussian noise with

$$E\{v(t_i)v(t_j)\} = R\,\delta_{ij}, \qquad R = 0.5 \quad \deg^2/hr^2$$

Note that declaring $v(\cdot, \cdot)$ to be a white sequence is really assuming that the correlation time of any noise corrupting the analog measuring device output is short compared to the sample period of the sampler. It is desired to process these measurements to obtain an optimal estimate of the gyro drift rate $x(t)$.

First generate the state differential equation. Since the Laplace domain transfer function of the shaping filter is:

$$x(s)/w(s) = \alpha/(s + \alpha)$$

we get, by cross multiplying

$$sx(s) + \alpha x(s) = \alpha w(s)$$

or, taking the inverse Laplace transform

$$\dot{x}(t) = -\alpha x(t) + \alpha w(t)$$

from which we can identify $\mathbf{F} = -\alpha = -1$, $\mathbf{G} = \alpha = 1$.

If we want to use (5-36) and (5-37), we will need the state transition matrix, which in this case is

$$\Phi(t_i, t_{i-1}) = \exp[-\alpha(t_i - t_{i-1})] = \exp[-1(0.25)] \cong 0.78$$

This could also be obtained through the inverse Laplace transform of $(s\mathbf{I} - \mathbf{F})^{-1}$.

It was assumed that $x(\cdot, \cdot)$ is a stationary process, so we must determine its steady state variance to serve as P_0. The general stochastic process covariance relation

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^T(t)$$

becomes

$$\dot{P}(t) = -2\alpha P(t) + \alpha^2 Q$$

so that the steady state value, evaluated by solving $\dot{P}(t) = 0$, is

$$P = \alpha Q/2 = 1 \quad \deg^2/hr^2$$

Thus, before any measurements are taken, we have the initial conditions of

$$\hat{x}(t_0) = \hat{x}_0 = 0 \qquad \text{(assumed)}$$
$$P(t_0) = P_0 = 1 \quad \deg^2/hr^2$$

At this point, the filter can be completely delineated. To propagate the estimate from sample time t_{i-1} to the next time t_i, (5-36) and (5-37) yield

$$\hat{x}(t_i^-) = \Phi(t_i, t_{i-1})\hat{x}(t_{i-1}^+) = 0.78\hat{x}(t_{i-1}^+)$$

$$P(t_i^-) = \Phi^2(t_i, t_{i-1})P(t_{i-1}^+) + \int_{t_{i-1}}^{t_i} \Phi^2(t_i, \tau)G^2 Q \, d\tau$$

$$= 0.78^2 P(t_{i-1}^+) + 2 \int_{t_{i-1}}^{t_i} \exp[-2(t_i - \tau)] \, d\tau$$

$$= 0.61 P(t_{i-1}^+) + 0.39$$

Note that these can also be interpreted as the Kalman filter using (5-51) and (5-52), based on the equivalent discrete-time model with $\mathbf{B}_d = 0$, $\mathbf{G}_d = 1$, and $\mathbf{Q}_d = 0.39$. This propagation can also be represented in differential equation form by using (5-46) and (5-47) to write

$$\dot{\hat{x}}(t/t_{i-1}) = F(t)\hat{x}(t/t_{i-1}) = -\hat{x}(t/t_{i-1})$$

$$\dot{P}(t/t_{i-1}) = 2F(t)P(t/t_{i-1}) + G^2(t)Q(t)$$

$$= -2P(t/t_{i-1}) + 2$$

which would be integrated forward from $\hat{x}(t_{i-1}/t_{i-1}) = \hat{x}(t_{i-1}^+)$, $P(t_{i-1}/t_{i-1}) = P(t_{i-1}^+)$ to time t_i. To update the estimate with the measurement z_i at time t_i, (5-38)–(5-40) yield

$$K(t_i) = \frac{P(t_i^-)H(t_i)}{H(t_i)P(t_i^-)H(t_i) + R(t_i)} = \frac{P(t_i^-)}{P(t_i^-) + 0.5}$$

$$\hat{x}(t_i^+) = \hat{x}(t_i^-) + K(t_i)[z_i - H(t_i)\hat{x}(t_i^-)]$$

$$= \hat{x}(t_i^-) + \frac{P(t_i^-)}{P(t_i^-) + 0.5}[z_i - \hat{x}(t_i^-)]$$

$$P(t_i^+) = P(t_i^-) - K(t_i)H(t_i)P(t_i^-)$$

$$= P(t_i^-) - \frac{P(t_i^-)^2}{P(t_i^-) + 0.5} = \frac{0.5P(t_i^-)}{P(t_i^-) + 0.5} \qquad \blacksquare$$

In the derivation of the Kalman filter algorithm, $\mathbf{P}(t_i^-)$ and $\mathbf{P}(t_i^+)$ were *defined* in (5-10) and (5-16) as conditional covariances of the state at time t_i, conditioned on $\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}$ and $\mathbf{Z}(t_i) = \mathbf{Z}_i$, respectively. However, the recursion relations (5-37), (5-38), and (5-40) for these matrices do *not* depend upon the particular sequence of realized measurements. The covariances are statistically related to the measurement history through the $\mathbf{H}(t_i)$ and $\mathbf{R}(t_i)$ sequences, but they are not functions of the specific measurement values. Whereas the state estimates (conditional means) are a function of measurement realizations, the error committed by such an estimate will be shown to be independent of $\mathbf{Z}(t_i)$. Consequently, one can precompute the time history of $\mathbf{P}(t_i^-)$, $\mathbf{P}(t_i^+)$, and $\mathbf{K}(t_i)$ before actual measurement numbers $\mathbf{z}_i$ are available: in fact, even before the measuring devices themselves are available. Since $\mathbf{P}(t_i^-)$ and $\mathbf{P}(t_i^+)$ are both conditional state covariances and state estimation error covariances, this precomputability allows early design tradeoffs of estimation accuracy versus measuring device precision [i.e., $\mathbf{R}(t_i)$ time history] to ensure cost-effective systems that meet specifications.
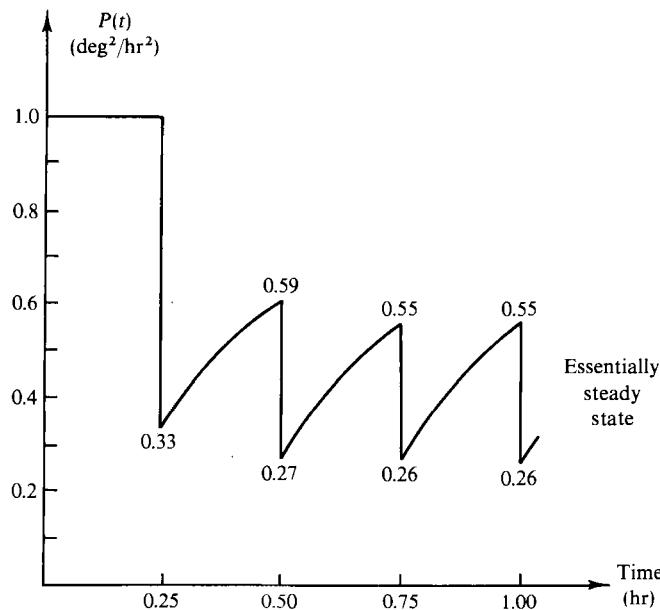
FIG. 5.6   Error variance time history for Example 5.3.

EXAMPLE 5.3   We can precompute the conditional state variance time history as generated by the Kalman filter for the problem of Example 5.2. Figure 5.6 portrays this variance time history through the fourth measurement sample time. [The values between sample times can be found from integrating (5-47), whereas (5-37) would generate only the $P(t_i^-)$ values.]

Note that the variance is constant over the first interval, since we started off in steady state conditions. Even on the first measurement, the estimate error variance is less than the measurement error variance of 0.5. Further, it decreases in time from 0.33 to 0.25: the algorithm uses the past history of measurement information, as propagated through its internal model, to yield this reduction.

Figure 5.6 indicates that steady state filter operation has been essentially achieved by the fourth measurement update time. To substantiate this claim, the steady state conditions can be calculated by equating $P(t_i^+)$ to $P(t_{i-1}^+)$:

$$P(t_i^+) = \frac{0.5P(t_i^-)}{P(t_i^-) + 0.5} = \frac{0.5[0.61P(t_{i-1}^+) + 0.39]}{[0.61P(t_{i-1}^+) + 0.39] + 0.5}$$

$$= \frac{0.30P(t_{i-1}^+) + 0.19}{0.61P(t_{i-1}^+) + 0.89}$$

Equating this to $P(t_{i-1}^+)$ yields

$$0.61P^{+2} + 0.59P^+ - 0.19 = 0$$

for which the positive solution is

$$P^+ = 0.255$$

Then, the steady state P^- is found from

$$P^- = 0.61P^+ + 0.39 = 0.546$$

These values do confirm the claim. ∎

The filter performance exhibited in the previous example is typical for problems with time-invariant system models and stationary statistics: an initial transient in $\mathbf{P}$ and $\mathbf{K}$ followed by an essentially steady state filter operation. In many applications, the transient is short compared to the total time of interest (suggesting a possible approximation of using the steady state filter for all time if the resulting performance degradation is not prohibitive; this will be discussed further in the next chapter). A different $\mathbf{P}_0$ matrix will yield a different magnitude transient characteristic, but its duration will be the same and the steady state conditions are unaffected.

On the other hand, changing $\mathbf{Q}$ or $\mathbf{R}$ in the filter structure does affect the transient duration and the steady state operation. Increasing $\mathbf{Q}$ would indicate either stronger noises driving the dynamics or increased uncertainty in the adequacy of the model itself to depict the true dynamics accurately. This will increase both the rate of growth of the $\mathbf{P}(t)$ elements (or eigenvalues) between measurement times and their steady state values. As a reult, the filter gains will generally increase, thereby weighting the measurements more heavily: this is reasonable since increased $\mathbf{Q}$ dictates that we should put less confidence in the output of the filter's own dynamics model. By similar reasoning, increased $\mathbf{R}$ would indicate that the measurements are subjected to a stronger corruptive noise, and so should be weighted less by the filter. In fact, this will decrease the gain values, the eigenvalues of the $[\mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i^-)]$ term are smaller so the error variances going from t_i^- to t_i^+ decrease to a lesser extent, and the steady state covariance eigenvalues are larger. If the eigenvalues of $\mathbf{Q}$ are large compared to the eigenvalues of $\mathbf{R}$ (in the scalar case, if the Q/R ratio is large), steady state is quickly reached because the uncertainty involved in the state propagation is large compared to the accuracy of the measurements, so the new state estimate is heavily dependent upon the new measurement and not closely related to prior estimates.

EXAMPLE 5.4   To see the effects of variations in Q and R, Example 5.3 will be repeated for the following cases (case 1 is Example 5.3 itself).

Case	Q (deg^2/hr)	R (deg^2/hr^2)
1	2	0.5
2	4	0.5
3	2	1.0
4	4	1.0

P_0 was set to the stationary value, which is 1 deg^2/hr^2 for $Q = 2$ deg^2/hr, and 2 deg^2/hr^2 for $Q = 4$ deg^2/hr; the same P_0 could have been used for all cases, but the differing transient would quickly decay and the same steady state filter operation would be achieved. Table 5.1 displays the time history of $P(t_i^-)$, $P(t_i^+)$, and $K(t_i)$ values for the four cases.

In this scalar example, the "tracking" properties of the filter will be very evident: if $K(t_i)$ is approximately one, $\hat{x}(t_i^+)$ is approximately equal to z_i. Conversely, if $K(t_i)$ is very small, then $\hat{x}(t_i)$ is not "tracking" the measurements closely, but rather is heavily weighting the output of its own internal system model.

TABLE 5.1

Effects of Q and R Variation

	Time (hr)	0.25	0.50	0.75	1.00
Case 1	$Q = 2, R = 0.5$				
	$P(t_i^-)$	1.00	0.59	0.55	0.55
	$P(t_i^+)$	0.33	0.27	0.26	0.26
	$K(t_i)$	0.67	0.54	0.52	0.52
Case 2	$Q = 4, R = 0.5$				
	$P(t_i^-)$	2.00	1.02	0.99	0.98
	$P(t_i^+)$	0.40	0.34	0.33	0.33
	$K(t_i)$	0.80	0.67	0.66	0.66
Case 3	$Q = 2, R = 1.0$				
	$P(t_i^-)$	1.00	0.69	0.64	0.63
	$P(t_i^+)$	0.50	0.41	0.39	0.39
	$K(t_i)$	0.50	0.41	0.39	0.39
Case 4	$Q = 4, R = 1.0$				
	$P(t_i^-)$	2.00	1.19	1.11	1.09
	$P(t_i^+)$	0.67	0.54	0.52	0.52
	$K(t_i)$	0.67	0.54	0.52	0.52

Doubling Q and retaining the original R (case 2) causes the difference between $P(t_{i-1}^+)$ and $P(t_i^-)$ to increase over that of case 1: in steady state, this is $(0.98 - 0.33) = 0.65$ versus $(0.54 - 0.26) = 0.28$. Not only are the oscillations larger due to more rapid growth (more rapid input of uncertainty) between sample times, but the steady state $P(t_i^-)$ and $P(t_i^+)$ are larger as well. With the same measurement precision but more uncertainty in the system model, there is less certainty in the estimate. The gains $K(t_i)$ are larger, and the filter "tracks" the measurements to a greater degree.

Doubling R and keeping the same Q as in case 1 (case 3) causes the difference between $P(t_i^-)$ and $P(t_i^+)$ to decrease: in steady state, $(0.63 - 0.39) = 0.24$ as opposed to 0.28. Steady state $P(t_i^-)$ and $P(t_i^+)$ are larger, since there is now more noise corruption in the measurements. Furthermore, there is now greater uncertainty in the measurements relative to the model uncertainties, so the $K(t_i)$'s are smaller, and the filter no longer "tracks" the measurements as closely.

Doubling both Q and R yields case 4. Here the gain time history is identical to that of case 1: because we have a scalar linear system description, the ratio Q/R is the determining factor of steady state gain. In fact, for this problem, steady state filter gain can be shown to be

$$K = 0.16\{\sqrt{[(Q/R) + 2]^2 + 12.5(Q/R)} - [(Q/R) + 2]\}$$

Thus, proportionately increasing Q and R had no effect on the $K(t_i)$ time history, while $P(t_i^-)$ and $P(t_i^+)$ were doubled (a linear system with doubling of all input magnitudes).

"Tuning" a general Kalman filter involves achieving "good" values of $\mathbf{P}_0$, $\mathbf{Q}$, and $\mathbf{R}$, good in the sense that the best estimation accuracy is obtained from a specified Kalman filter structure. The task of determining the best set of matrices is considerably more difficult than the scalar case, but this example does provide some basic insights. ∎

5.4 STATISTICS OF PROCESSES WITHIN THE FILTER STRUCTURE

The previous section derived the Kalman filter, through which the conditional probability density $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$ could be generated explicitly for all time. This density function, or the mean and covariance which define it, provides all the possible information obtainable about the system *state*. Now we investigate the statistical description of some of the processes within the filter structure itself.

The *error* committed by using $\hat{\mathbf{x}}(t_i^+)$ as an estimator of $\mathbf{x}(t_i)$ would be defined as

$$\mathbf{e}(t_i^+) \triangleq \mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^+) \tag{5-53}$$

where $\hat{\mathbf{x}}(t_i^+)$ is the random variable $E\{\mathbf{x}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}(t_i,\cdot)\}$, one realization of which would be the numerical output of the filter algorithm. Since $\mathbf{x}(t_i)$ and $\hat{\mathbf{x}}(t_i^+)$ can be shown to be jointly Gaussian, conditioned on $\mathbf{Z}(t_i)$, $\mathbf{e}(t_i^+)$ is a Gaussian random variable, and its density function is then completely specified by its mean and covariance. The mean is

$$E\{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i) = \mathbf{Z}_i\} = E\{\mathbf{x}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}_i\} - E\{\hat{\mathbf{x}}(t_i^+)|\mathbf{Z}(t_i) = \mathbf{Z}_i\}$$
$$= \hat{\mathbf{x}}(t_i^+) - \hat{\mathbf{x}}(t_i^+) = \mathbf{0} \tag{5-54}$$

i.e., the estimator is unbiased since the error is zero mean. The covariance is:

$$E\{\mathbf{e}(t_i^+)\mathbf{e}^{\mathrm{T}}(t_i^+)|\mathbf{Z}(t_i) = \mathbf{Z}_i\}$$
$$= E\{[\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^+)][\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^+)]^{\mathrm{T}}|\mathbf{Z}(t_i) = \mathbf{Z}_i\}$$
$$= E\{\mathbf{x}(t_i)\mathbf{x}^{\mathrm{T}}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}_i\} - E\{\mathbf{x}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}_i\}\hat{\mathbf{x}}^{\mathrm{T}}(t_i^+)$$
$$\quad - \hat{\mathbf{x}}(t_i^+)E\{\mathbf{x}^{\mathrm{T}}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}_i\} + \hat{\mathbf{x}}(t_i^+)\hat{\mathbf{x}}^{\mathrm{T}}(t_i^+)$$
$$= [\mathbf{P}(t_i^+) + \hat{\mathbf{x}}(t_i^+)\hat{\mathbf{x}}^{\mathrm{T}}(t_i^+)] - \hat{\mathbf{x}}(t_i^+)\hat{\mathbf{x}}^{\mathrm{T}}(t_i^+) - \hat{\mathbf{x}}(t_i^+)\hat{\mathbf{x}}^{\mathrm{T}}(t_i^+) + \hat{\mathbf{x}}(t_i^+)\hat{\mathbf{x}}^{\mathrm{T}}(t_i^+)$$
$$= \mathbf{P}(t_i^+) \tag{5-55}$$

This proves the previous statement that the conditional covariance of the error committed by using $\hat{\mathbf{x}}(t_i^+)$ as an estimator is equal to the conditional covariance of $\mathbf{x}(t_i)$ itself. Consequently, we can write

$$f_{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i) = [(2\pi)^{n/2}|\mathbf{P}(t_i^+)|^{1/2}]^{-1}\exp\{-\tfrac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{P}(t_i^+)^{-1}\boldsymbol{\xi}\} \tag{5-56}$$

This is functionally independent of the particular realized measurement values, $\mathbf{Z}_i$, since $\mathbf{P}(t_i^+)$ does not depend on $\mathbf{Z}_i$; for this reason, it is often written as $f_{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i)}(\boldsymbol{\xi})$. By exploiting this and the concepts of marginal densities and Bayes' rule, the unconditional density $f_{\mathbf{e}(t_i^+)}(\boldsymbol{\xi})$ can be written as

$$f_{\mathbf{e}(t_i^+)}(\boldsymbol{\xi}) = \int_{-\infty}^{\infty} f_{\mathbf{e}(t_i^+),\,\mathbf{Z}(t_i)}(\boldsymbol{\xi}, \mathscr{Z}_i)\,d\mathscr{Z}_i = \int_{-\infty}^{\infty} f_{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i)}(\boldsymbol{\xi})\,f_{\mathbf{Z}(t_i)}(\mathscr{Z}_i)\,d\mathscr{Z}_i$$

$$= f_{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}) \int_{-\infty}^{\infty} f_{\mathbf{Z}(t_i)}(\mathscr{Z}_i)\,d\mathscr{Z}_i = f_{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}) \cdot 1 \qquad (5\text{-}57)$$

Thus, if we use $\hat{\mathbf{x}}(t_i^+)$ as our state estimator, the error committed, $\mathbf{e}(t_i^+)$, is *independent* of $\mathbf{Z}(t_i)$, the entire measurement history random variable.

Figure 5.7 portrays this graphically. If the time history of measurements $\mathbf{Z}_i$ changes, then the conditional mean $\hat{\mathbf{x}}(t_i^+)$ changes, but the shape of the density function $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$ [which is just $f_{\mathbf{e}(t_i^+)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$ shifted by the conditional mean] remains unchanged. This is a unique characteristic, not true of nonlinear estimation problems in general.

This independence says conceptually that the estimator gleans out as much information from the measurements as possible, and there is nothing left in the measurements that could tell you anything about the error. Geometrically, the
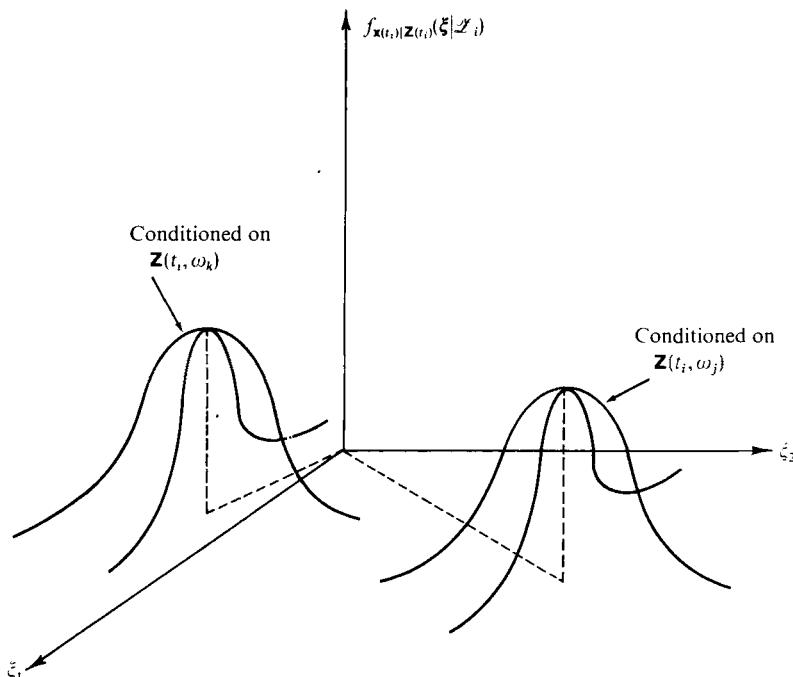
FIG. 5.7 Estimation error independence of measurement history.

error is orthogonal to the projection of the real $\mathbf{x}(t_i)$ onto the measurement subspace (in a Hilbert space of random variables): Kalman originally developed the filter recursion relations from this geometrical insight.

Similarly, if we define

$$\mathbf{e}(t_i^-) \triangleq \mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^-) \tag{5-58}$$

then, $f_{\mathbf{e}(t_i^-)|\mathbf{Z}(t_{i-1})}(\boldsymbol{\xi}\,|\,\mathbf{Z}_{i-1})$ can be shown to be Gaussian, with

$$E\{\mathbf{e}(t_i^-)\,|\,\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} = \mathbf{0} \tag{5-59}$$

$$E\{\mathbf{e}(t_i^-)\mathbf{e}^{\mathrm{T}}(t_i^-)\,|\,\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} = \mathbf{P}(t_i^-) \tag{5-60}$$

Two other closely related processes in the filter are the *residual* (*innovations*) and *new information* processes, denoted as $\mathbf{r}(\cdot,\cdot)$ and $\mathbf{s}(\cdot,\cdot)$, respectively, and defined for all $t_i \in T$ by

$$\mathbf{r}(t_i) \triangleq \mathbf{z}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) \tag{5-61}$$

$$\mathbf{s}(t_i) \triangleq \mathbf{K}(t_i)\mathbf{r}(t_i) \tag{5-62}$$

The term $[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$ in the filter measurement update equation (5-39) is a realization of $\mathbf{r}(t_i)$: the difference between the current measurement value $\mathbf{z}_i$ and the best prediction of its value before the measurement is actually taken. It is then a particular realization of $\mathbf{s}(t_i)$ that is added to $\hat{\mathbf{x}}(t_i^-)$ to obtain $\hat{\mathbf{x}}(t_i^+)$. We can rewrite $\mathbf{r}(t_i)$ as

$$\begin{aligned}
\mathbf{r}(t_i) &= \mathbf{z}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) \\
&= \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) \\
&= \mathbf{H}(t_i)\left[\boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \right.\\
&\quad \left. + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}(\tau)\,d\boldsymbol{\beta}(\tau) \right] + \mathbf{v}(t_i) \\
&\quad - \mathbf{H}(t_i)\left[\boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+) + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \right] \\
&= \mathbf{H}(t_i)\boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{e}(t_{i-1}^+) + \mathbf{H}(t_i)\int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}(\tau)\,d\boldsymbol{\beta}(\tau) + \mathbf{v}(t_i) \tag{5-63}
\end{aligned}$$

But $\mathbf{e}(t_{i-1}^+)$, $\int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}(\tau)\,d\boldsymbol{\beta}(\tau)$, and $\mathbf{v}(t_i)$ are all random variables that are independent of $\mathbf{Z}(t_{i-1})$, so $\mathbf{r}(t_i)$ is *independent of* $\mathbf{Z}(t_{i-1})$. But, by their definition, $\mathbf{r}(t_1)$, $\mathbf{r}(t_2)$, ..., $\mathbf{r}(t_{i-1})$ are linear functions of $\mathbf{Z}(t_{i-1})$, so $\mathbf{r}(t_i)$ is independent of all previous $\mathbf{r}(t_j)$'s. In other words, *the* $\mathbf{r}(t_i)$ *sequence is a white sequence*. In view of (5-61), this also demonstrates that $\mathbf{s}(t_i)$ *is independent of* $\mathbf{Z}(t_{i-1})$ and that the $\mathbf{s}(t_i)$ *sequence is white*: each new piece of information is independent of the information gained in the past or, geometrically, $\mathbf{s}(t_i)$ is orthogonal to $\hat{\mathbf{x}}(t_i^-)$.

Moreover, based on arguments of linear combinations of jointly Gaussian random variables being Gaussian, $\mathbf{r}(t_i)$ and $\mathbf{s}(t_i)$ are both *Gaussian* for all t_i. To describe these processes completely, we only need to specify means and covariance kernels (conditioned on measurements or not does not matter, but quantities *defined* as conditional entities will be identified, so we choose to condition on $\mathbf{Z}_{i-1}$):

$$E\{\mathbf{r}(t_i)\} = E\{\mathbf{r}(t_i)|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} = \mathbf{0} \qquad (5\text{-}64a)$$

$$\begin{aligned} E\{\mathbf{r}(t_i)\mathbf{r}^{\mathsf{T}}(t_i)\} &= E\{\mathbf{r}(t_i)\mathbf{r}^{\mathsf{T}}(t_i)|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \\ &= E\{(\mathbf{H}(t_i)[\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^-)] + \mathbf{v}(t_i)) \\ &\quad \cdot (\mathbf{H}(t_i)[\mathbf{x}(t_i) - \hat{\mathbf{x}}(t_i^-)] + \mathbf{v}(t_i))^{\mathsf{T}}|\mathbf{Z}(t_{i-1}) = \mathbf{Z}_{i-1}\} \\ &= \mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i) \qquad (5\text{-}64b) \end{aligned}$$

$$E\{\mathbf{s}(t_i)\} = \mathbf{K}(t_i)E\{\mathbf{r}(t_i)\} = \mathbf{0} \qquad (5\text{-}65a)$$

$$\begin{aligned} E\{\mathbf{s}(t_i)\mathbf{s}^{\mathsf{T}}(t_i)\} &= \mathbf{K}(t_i)E\{\mathbf{r}(t_i)\mathbf{r}^{\mathsf{T}}(t_i)\}\mathbf{K}^{\mathsf{T}}(t_i) \\ &= \mathbf{K}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i)]\mathbf{K}^{\mathsf{T}}(t_i) \\ &= \mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i)]^{-1}\mathbf{H}(t_i)\mathbf{P}(t_i^-) \quad (5\text{-}65b) \end{aligned}$$

Obtaining (5-65b) exploited the symmetry of $[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i)]$; the result is singular in general, being n-by-n and of rank at most m. As expected, we recognize (5-65b) as the term that is subtracted from $\mathbf{P}(t_i^-)$ to obtain $\mathbf{P}(t_i^+)$: the decrease in estimation error covariance due to incorporating the information of the measurement at time t_i.

The residual sequence has been shown to be a white Gaussian sequence of mean zero and covariance $[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i)]$. This can be exploited for the practical purposes of either sensor failure detection or reasonableness checking of measurement data. The preceding is a proper characterization of the residual process provided that the mathematical model upon which the filter was based accurately depicts the real system behavior. During operation of the filter, the actual residual sequence realization can be monitored and compared to this description. If the description appears valid up to a point in time, thereafter being violated consistently, one can deduce that something occurred in the real system to invalidate the model within the filter. If the violation occurs in only one component of a vector residual process, one can further deduce that the measuring device generating that particular residual component is the source of difficulty: a failure (soft or hard) can be declared in that sensor.

Optimal "likelihood function" methods or ad hoc techniques of event detection (hypothesis testing) can be used to perform a test for the occurrence of a sensor failure [51,52]. Essentially, the N most recent residual signals are examined to determine whether they differ significantly from the statistical description of their values that assumes no failures [9, 55]. The number N is a design parameter. It is kept greater than one to prevent failure declarations due

to a single residual sample of large magnitude: consistently large residuals indicate abnormalities, whereas individual realizations of large magnitude are to be expected. On the other hand, it is inappropriate to use all the residual samples from the initial time to current time, since this would decrease the sensitivity to true failures as time progressed. Thus, a "moving window" of the N most recent samples, with N on the order of 5 to 20, would be used.

Statistical hypothesis testing theory indicates that a good choice of likelihood function [67] for event (failure) detection would be in the form of sum of natural logs of conditional densities for components of residuals: for the kth component,

$$L_{N_k}(t_i) = \sum_{j=i-N+1}^{i} \ln f_{r_k(t_j)|r_k(t_{j-1}),\ldots,r_k(t_1)}(\rho_j|\rho_{j-1},\ldots,\rho_1) \qquad (5\text{-}66)$$

If the residual sequence can be assumed to be a set of independent zero-mean Gaussian random variables, then this can be rewritten as

$$L_{N_k}(t_i) = c_k(t_i) - \frac{1}{2}\sum_{j=i-N+1}^{i} \frac{r_k^2(t_j)}{\sigma_k^2(t_j)} \qquad (5\text{-}67)$$

where $c_k(t_i)$ is a (slowly varying) negative term independent of the observed residual values (thus containing no information of direct use for failure detection), and $\sigma_k^2(t_j)$ is the estimate of the variance of possible kth residual values based on the assumption that no failures have occurred. The value of $1/\sigma_k^2(t_j)$ can be evaluated as the kth diagonal term of $[\mathbf{H}(t_j)\mathbf{P}(t_j^-)\mathbf{H}^T(t_j) + \mathbf{R}(t_j)]^{-1}$, a matrix that has *already* been computed in the filter algorithm. If $r_k^2(t_j)$ becomes consistently larger than that predicted by $\sigma_k^2(t_j)$ over the most recent N samples, $L_{N_k}(t_i)$ will become more and more negative; if its value goes beyond a predetermined threshold, a failure can be declared.

EXAMPLE 5.5   Figure 5.8a portrays a possible residual process realization in the filter described in Example 5.2. Until time t_F (time of sensor failure), the residual sequence is well described as a zero-mean white Gaussian sequence of variance $\sigma^2 = [\mathbf{HPH}^T + R]$: about 68°ₒ of the samples lie within the 1σ bounds, 95% within the 2σ bounds, etc. At t_F, the bias shifts markedly from zero. In Fig. 5.8b, the residual process strength increases markedly at time t_F. For either case, the likelihood function $L_N(t_i)$ would be as depicted in Fig. 5.8c: from time t_F on, the $r^2(t_j)/\sigma^2(t_j)$ terms are larger, so $L_N(t_i)$ grows more negative. Upon passing the threshold, a failure is declared.   ■

Residual monitoring is also used for reasonableness checking of measurements before they are processed by the filter. If a spurious data point is received, we would want to reject it rather than let it corrupt the filter computations. If a measurement residual is greater than (for instance) the 3σ value computed by the filter, it can be declared as unacceptable. However, if this happens on a frequent basis, the cause may be either a sensor failure or a filter divergence problem (the filter's estimates do not correspond well to the real system behavior). In this latter case, it is critical not to reject large residuals, since they are the only means of correcting the divergence! Filter divergence [25] and more sophisticated use of residual monitoring (adaptive estimation algorithms,
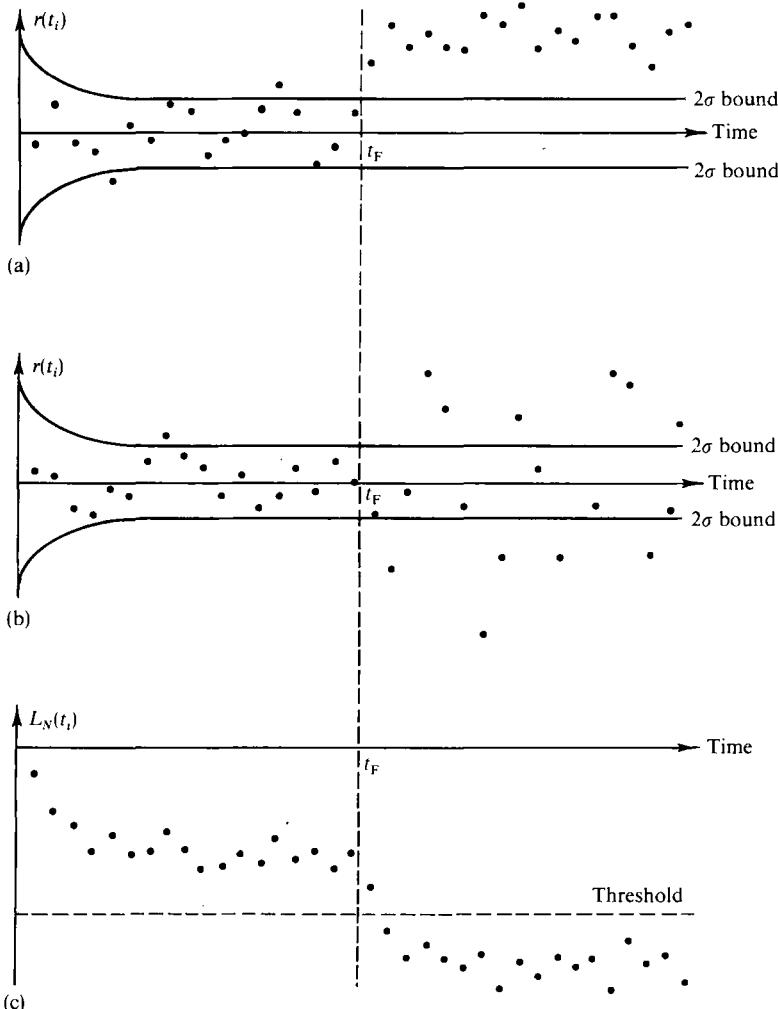
FIG. 5.8 Residual monitoring and sensor failure detection. (a) Residual bias shift. (b) Residual strength increase. (c) Likelihood function $L_N(t_i)$ for either case.

which modify their internal models and/or gains online by exploiting the observed residuals) will be discussed in detail in subsequent chapters.

5.5 OTHER CRITERIA OF OPTIMALITY

In Section 5.3 we derived the Kalman filter algorithm in a Bayesian manner by generating explicit recursions for the Gaussian conditional probability density for the states, conditioned on the entire measurement history, $f_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi} \,|\, \mathbf{Z}_i)$. Then $\hat{\mathbf{x}}(t_i^+)$ was chosen as the "optimal estimate" because it was

the mean, mode, and median of this density function. There are other criteria of optimality that are logical for an estimation problem [16, 18, 29, 31, 64, 68, 82], and thus there are other means of deriving the filter relations. Some of these aspects will now be discussed.

By virtue of being the conditional mean, $\hat{\mathbf{x}}(t_i^+)$ *is also the minimum mean square error* (MMSE) *estimate* [54, 82]. In a general estimation problem, if $\hat{\mathbf{x}}_{\text{EST}}(t_i)$ is some estimator of $\mathbf{x}(t_i)$ and $\mathbf{e}_{\text{EST}}(t_i)$ is the error committed by this estimator,

$$\mathbf{e}_{\text{EST}}(t_i) = \mathbf{x}(t_i) - \hat{\mathbf{x}}_{\text{EST}}(t_i) \tag{5-68}$$

then the estimator that minimizes the cost function

$$J[\hat{\mathbf{x}}_{\text{EST}}(t_i)] = E\{\mathbf{e}_{\text{EST}}(t_i)^{\text{T}}\mathbf{e}_{\text{EST}}(t_i)\} \tag{5-69}$$

is the conditional mean. This is true for any form of probability distribution function $F_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$. Moreover, if $F_{\mathbf{x}(t_i)|\mathbf{Z}(t_i)}(\boldsymbol{\xi}|\mathbf{Z}_i)$ is Gaussian, as is true for the particular problem addressed in this chapter, the conditional mean also minimizes *any* cost function of the general quadratic form

$$J[\hat{\mathbf{x}}_{\text{EST}}(t_i)] = E\{\mathbf{e}_{\text{EST}}(t_i)^{\text{T}}\mathbf{M}(t_i)\mathbf{e}_{\text{EST}}(t_i)\} \tag{5-70}$$

where $\mathbf{M}(t_i)$ for all $t_i \in T$ form a set of arbitrary symmetric, positive semi-definite matrices. For further discussion of *least squares estimation*, see [1, 5, 24, 27, 28, 32, 33, 57, 64–66, 73, 75–77].

EXAMPLE 5.6   Consider an estimation problem with a two dimensional state vector. Consider $\mathbf{M}(t_i)$ constant in time, set equal to any of the following:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$$

According to the above claim, the *same* estimator (the conditional mean) would optimize (5-70) for all five choices. Whether you are interested in estimating only x_1, only x_2, or both with any relative importance in estimation accuracy, the same estimator would be used.   ∎

Also due to its being the conditional mean, $\hat{\mathbf{x}}(t_i^+)$ *minimizes the symmetric cost function criterion* [14, 54, 69, 70, 82]. Define a general estimation error cost function as

$$J[\hat{\mathbf{x}}_{\text{EST}}(t_i)] = E\{C[\mathbf{e}_{\text{EST}}(t_i)]\} \tag{5-71}$$

If $C(\cdot)$ is symmetric and nondecreasing,

$$\begin{aligned} C(\mathbf{0}) &= 0 \\ C(\mathbf{e}) &= C(-\mathbf{e}) \\ C(\mathbf{e}_2) &\geq C(\mathbf{e}_1) \quad \text{if} \quad \|\mathbf{e}_2\| > \|\mathbf{e}_1\| \end{aligned} \tag{5-72}$$

and if $f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi\,|\,\mathbf{Z}_i)$ is unimodal (one-peaked), symmetric about the conditional mean, and satisfies

$$\lim_{||\xi||\to\infty} C(\xi)f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi\,|\,\mathbf{Z}_i) = 0 \tag{5-73}$$

then the cost function (5-71) is minimized by the conditional mean. For the problem at hand, the assumptions on $f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi\,|\,\mathbf{Z}_i)$ are met, and we need only assume (5-72) to be true. Note particularly that there was no need to assume $C(\cdot)$ to be convex, so that all the cost functions depicted in Fig. 5.9 are admissible. Comparing this claim to the fact that the conditional mean is the MMSE estimate, optimality with respect to a more general cost criterion has
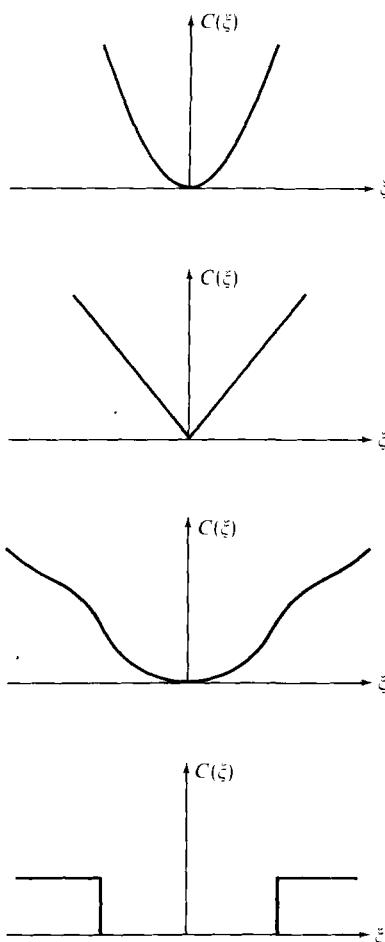
FIG. 5.9   Cost functions admissible according to Eq. (5-72).

been achieved, but at the expense of additional restrictions on the allowable class of stochastic processes [the assumptions on $f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi|\mathbf{Z}_i)$].

To obtain a *maximum likelihood estimate* of the system state, an appropriate "likelihood function" [17, 67, 84] must be defined as a scalar function relating the available measurements (whose values are known), the state variables (the unknowns to be estimated), and any other pertinent parameters. One choice of a likelihood function (though not the "classical" choice) would be $f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi|\mathbf{Z}_i)$ itself. By maximizing this likelihood function, we are actually finding the mode (location of the peak) of the conditional density, and the resulting estimator is often called the maximum a posteriori, or MAP, estimate.

To show that $\hat{\mathbf{x}}(t_i^+)$ *is the* MAP *estimate of the state* [30, 62, 66, 82] involves an algebraically simpler derivation than the reduction that followed Eq. (5-27) originally. Under the assumptions of our problem formulation, $f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi|\mathcal{Z}_i)$ is Gaussian, so it is more convenient to define the (log-) likelihood function as the natural logarithm of the conditional density,

$$L(\xi, \mathcal{Z}_i) = \ln f_{\mathbf{x}(t_i)|\mathbf{z}(t_i)}(\xi|\mathcal{Z}_i) \tag{5-74}$$

Maximizing this function yields the same estimate as maximizing the density itself, since for any function f, f and $\ln f$ attain their maxima at the same point. Substitute (5-27) into (5-74) to obtain

$$
\begin{aligned}
L(\xi, \mathcal{Z}_i) = \ln\{ &(2\pi)^{-n/2} |\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)|^{1/2} \\
&\times |\mathbf{P}(t_i^-)|^{-1/2} |\mathbf{R}(t_i)|^{-1/2} \} \\
&- \tfrac{1}{2}\{ [\zeta_i - \mathbf{H}(t_i)\xi]^T \mathbf{R}(t_i)^{-1} [\zeta_i - \mathbf{H}(t_i)\xi] \} \\
&- \tfrac{1}{2}\{ [\xi - \hat{\mathbf{x}}(t_i^-)]^T \mathbf{P}(t_i^-)^{-1} [\xi - \hat{\mathbf{x}}(t_i^-)] \} \\
&+ \tfrac{1}{2}\{ [\zeta_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]^T [\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)]^{-1} \\
&\times [\zeta_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \}
\end{aligned}
\tag{5-75}
$$

To generate the MAP estimate of $\mathbf{x}(t_i)$, denoted as $\hat{\mathbf{x}}_{\text{MAP}}(t_i)$, we must solve

$$\partial L[\xi, \mathcal{Z}_i]/\partial \xi \big|_{\xi \to \hat{\mathbf{x}}_{\text{MAP}}(t_i)} = \mathbf{0}^T \tag{5-76}$$

Performing this differentiation on (5-75) yields

$$\{\mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)[\zeta_i - \mathbf{H}(t_i)\xi] - \mathbf{P}(t_i^-)^{-1}[\xi - \hat{\mathbf{x}}(t_i^-)]\}\big|_{\xi \to \hat{\mathbf{x}}_{\text{MAP}}(t_i)} = \mathbf{0}$$

or

$$\{\mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)\xi + \mathbf{P}(t_i^-)^{-1}\xi\}\big|_{\xi \to \hat{\mathbf{x}}_{\text{MAP}}(t_i)} = \mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\zeta_i + \mathbf{P}(t_i^-)^{-1}\hat{\mathbf{x}}(t_i^-)$$

for which the solution is

$$\hat{\mathbf{x}}_{\text{MAP}}(t_i) = [\mathbf{P}(t_i^-)^{-1} + \mathbf{H}^{\text{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)]^{-1}[\mathbf{H}^{\text{T}}(t_i)\mathbf{R}^{-1}(t_i)\boldsymbol{\zeta}_i + \mathbf{P}(t_i^-)^{-1}\hat{\mathbf{x}}(t_i^-)]$$

$$(5\text{-}77)$$

Comparing this to (5-31) reveals that $\hat{\mathbf{x}}_{\text{MAP}}(t_i) = \hat{\mathbf{x}}(t_i^+)$.

The classical maximum likelihood estimate (often denoted as MLE) is found by maximizing the likelihood function chosen to be the conditional density $f_{\mathbf{z}(t_i)|\mathbf{x}(t_i)}(\mathscr{Z}_i|\boldsymbol{\xi})$, or its natural logarithm. Conceptually, maximizing this maximizes the probability of the event that *did* in fact occur, i.e., $\mathbf{Z}(t_i, \omega_j) = \mathscr{Z}_i$, expressed as a function of $\boldsymbol{\xi}$. It can be shown that, under the assumptions of the original derivation, the value of $\boldsymbol{\xi}$ which maximizes this likelihood function is again $\hat{\mathbf{x}}(t_i^+)$ if there is no a priori state information (an MLE does not incorporate such data): $\hat{\mathbf{x}}(t_i^+)$ *is the maximum likelihood estimate* (MLE) *of* $\mathbf{x}(t_i)$ *if* $\mathbf{P}_0 = \infty\mathbf{I}$, i.e. *if* $\mathbf{P}_0^{-1} = \mathbf{0}$, *and converges asymptotically to the* MLE *if* $\mathbf{P}_0^{-1} \neq \mathbf{0}$ [60].

As a maximum likelihood estimator, $\hat{\mathbf{x}}(t_i^+)$ possesses certain desirable characteristics [17, 23, 49, 82, 84]. Under rather general regularity conditions, a general maximum likelihood parameter estimator can be shown to be consistent (it converges to the true value as the number of measurement samples grows without bound), asymptotically unbiased, asymptotically normally distributed, and asymptotically efficient (as the number of samples grows without bound, it is unbiased, has finite covariance, and there is no other unbiased estimate whose covariance is smaller). For this particular problem formulation, these asymptotic properties are also true for a finite number of measurement samples.

Furthermore, $\hat{\mathbf{x}}(t_i^+)$ *is the minimum variance unbiased linear estimate* [2–5, 26, 66, 72] of the state. Consider the same problem formulation as in Section 5.2, except that the noises need not be Gaussian. Then $\hat{\mathbf{x}}(t_i^+)$ is the estimator out of the class of *linear* unbiased (zero-mean error) estimators that yields a minimum error variance (minimum trace of the error covariance matrix). It is the best linear estimator in this sense, but there are nonlinear estimators which may outperform it. In the original derivation, we imposed the Gaussian assumption and did *not* have to seek the best *linear* filter: under this additional assumption, the linear filter *is* the best filter of any kind.

The original derivation of the Kalman filter was based on the fact that $\hat{\mathbf{x}}(t_i^+)$ *is the orthogonal projection of the true state* $\mathbf{x}(t_i)$ *onto the subspace spanned by* $\mathbf{Z}(t_i)$ [35, 47, 64]. The orthogonal projection lemma of functional analysis is applied to the estimation problem as posed in the infinite-dimensional Hilbert space of random variables with finite second moments. We will not delve into the rigor of the proof, but the geometric insight is that if the estimate of $\mathbf{x}(t_i) \in \mathscr{X}$ is desired in the form of some $\hat{\mathbf{x}}_{\text{EST}}(t_i)$ confined to a subspace of $\mathscr{X}$, then the

best estimate (approximation) is the orthogonal projection of $\mathbf{x}(t_i)$ onto that subspace: such that the error defined in (5-68) is orthogonal to that subspace.

Thus, the Kalman filter algorithm is optimal with respect to many different criteria. This reveals the power and importance of this algorithm, the practical design and implementation of which will be detailed subsequently.

5.6 COVARIANCE MEASUREMENT UPDATE COMPUTATIONS

The most troublesome numerical aspect of the Kalman filter is the measurement update of the covariance matrix, so the properties of alternate computational forms are of substantial interest.

As originally obtained in the derivation of the estimator, this update can be written as

$$\mathbf{P}(t_i{}^+) = \left[\mathbf{P}(t_i{}^-)^{-1} + \mathbf{H}^{\mathrm{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i) \right]^{-1} \qquad (5\text{-}78)$$

Although this form adds the measurement information in a simple manner that preserves symmetry well, it requires two n-by-n inversions each sample time (unless an inverse covariance is used in place of the covariance, as discussed in the next section), where n is the number of state variables.

Applying the matrix inversion lemma to (5-78) provided an alternate form of

$$\mathbf{P}(t_i{}^+) = \mathbf{P}(t_i{}^-) - \mathbf{P}(t_i{}^-)\mathbf{H}^{\mathrm{T}}(t_i)\left[\mathbf{H}(t_i)\mathbf{P}(t_i{}^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i)\right]^{-1}\mathbf{H}(t_i)\mathbf{P}(t_i) \qquad (5\text{-}79\mathrm{a})$$

$$= \mathbf{P}(t_i{}^-) - \mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i{}^-) \qquad (5\text{-}79\mathrm{b})$$

This involves m-by-m inversions, where m is the dimension of the measurement vector. For m significantly less than n, as is the case in many practical applications, (5-79) is therefore much more efficient than (5-78). However, (5-79) can involve the small difference of large numbers, especially if the measurements are very accurate. On a finite word length computer, this can cause serious numerical precision problems, even to the extent of not assuring positive definiteness of the result. (Once any computed covariance obtains a negative eigenvalue, all subsequent computations are erroneous since they are based on a theoretical impossibility. In practice, filters can be "tuned" or modified so as to be able to avoid or recover from such numerical errors, as will be seen in subsequent chapters.)

EXAMPLE 5.7 Consider the first measurement time in the gyro on test introduced in Example 5.2, but let R be changed from 5×10^{-1} deg^2/hr^2 to 5×10^{-4} deg^2/hr^2. Then $P(t_i{}^-) = 1$, $K(t_i)H(t_i)P(t_i{}^-) = 0.99950025$, and $P(t_i{}^+) = 0.00049975$ to eight significant figures. To three significant figures, $K(t_i)H(t_i)P(t_i{}^-)$ would be rounded up to one and $P(t_i{}^+)$ would be zero. If truncation were used rather than rounding, then $K(t_i)H(t_i)P(t_i{}^-)$ would be 0.999 and $P(t_i{}^+)$ would be 0.001. Both cases are seen to be very erroneous on a percentage basis. ∎

The update (5-79b) is readily seen to be equivalent algebraically to

$$\mathbf{P}(t_i{}^+) = [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\mathbf{P}(t_i{}^-) \tag{5-80}$$

Not only does this form fail to assure positive definiteness as true of (5-79b), but it suffers additionally from the fact that symmetry is not well preserved either. Whereas (5-79b) entailed subtracting one symmetric form from another, (5-80) is in the form of a product of a nonsymmetric matrix and a symmetric one, and thus it is a less desirable form.

If the state estimate update equation is rewritten as

$$\hat{\mathbf{x}}(t_i{}^+) = [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\hat{\mathbf{x}}(t_i{}^-) + \mathbf{K}(t_i)\mathbf{z}_i \tag{5-81}$$

it can be readily shown that an equivalent expression for $\mathbf{P}(t_i{}^+)$ is the "Joseph form" (after the man who first developed it):

$$\mathbf{P}(t_i{}^+) = [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\mathbf{P}(t_i{}^-)[\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]^{\mathrm{T}} + \mathbf{K}(t_i)\mathbf{R}(t_i)\mathbf{K}^{\mathrm{T}}(t_i) \tag{5-82}$$

This is in the form of the *sum* of two *symmetric* matrices, the first being positive definite and the second being positive semidefinite (n-by-n, and of rank at most m). Consequently, numerical computations based upon this form will be better conditioned, better assuring both the symmetry and positive definiteness of $\mathbf{P}(t_i{}^+)$ than previous forms. Furthermore, it is insensitive, to first order, to small errors $\delta\mathbf{K}(t_i)$ in the computed filter gain: for a first order error $\delta\mathbf{K}(t_i)$, the error in the $\mathbf{P}(t_i{}^+)$ computed by (5-82) is of second order, while the error in the previous forms is of first order,

$$\delta\mathbf{P}(t_i{}^+) = -\delta\mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i{}^-) \tag{5-83}$$

Similarly, it is less sensitive to arithmetic truncation than the other forms: especially in the cases in which the measurement noise is small, (5-79) and (5-80) will be subject to first order truncation error effects, while (5-82) will only be affected to second order. This becomes a crucial consideration for online applications in which the minimum computer wordlength that achieves adequate performance is sought.

Although the Joseph form has some desirable characteristics, it requires a considerably greater number of computations (multiplications and additions), so more computer time is required. (Section 7.8 will tabulate required operations for various forms; Problem 5.17 develops a more efficient implementation of the Joseph form.) A tradeoff must be analyzed to determine if the benefits warrant the additional loading. In fact, there are some cases, especially those characterized by long periods of essentially steady state behavior, in which the inherently greater number of adds and multiplies (each individual operation with truncation or roundoff effects of its own) causes larger numerical errors in the Joseph form than in the others.

To date, it is typical to perform the majority of the algorithm computations in single precision. Regardless of the form used, the covariance measurement update calculations are done in double precision to maintain numerical accuracy.

For online applications in which time constraints are critical, symmetry can be exploited by propagating and updating only lower triangular forms of the covariance matrix. This requires only $\frac{1}{2}n(n + 1)$ scalar terms instead of n^2, which can be substantial for large n. However, due to symmetry preservation problems, some operational filters maintain all n^2 terms and periodically resymmetrize the covariance matrix by averaging the appropriate elements.

Symmetry can be exploited further by using a square root covariance formulation: express and compute the algorithm results in terms of $\mathbf{P}^{1/2}$ instead of $\mathbf{P}$ (a matrix $\mathbf{P}^{1/2}$ such that $\mathbf{P}^{1/2}\mathbf{P}^{1/2\,\mathrm{T}} = \mathbf{P}$ can always be defined for a symmetric positive definite $\mathbf{P}$ matrix). This attains equivalent numerical accuracy with approximately half the wordlength, while requiring a number of computations comparable to that of the Joseph form update for $\mathbf{P}$. A related technique, known as $\mathbf{U}$–$\mathbf{D}$ covariance factorization, in which $\mathbf{P}$ is factored as $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, with $\mathbf{U}$ being upper triangular and unitary and $\mathbf{D}$ being diagonal, provides the same numerical benefits but with considerably less computational loading. "Square root filtering" is the subject of Chapter 7.

An alternate expression for the filter gain $\mathbf{K}(t_i)$, given originally by (5-38), is

$$\mathbf{K}(t_i) = \mathbf{P}(t_i^{+})\mathbf{H}^{\mathrm{T}}(t_i)\mathbf{R}^{-1}(t_i^{-}) \tag{5-84}$$

Equivalence can be demonstrated by substituting (5-79a) into (5-84) to obtain (5-38). Although (5-84) is a simpler expression, it is not very useful for the discrete-time update (sampled-data) filter formulation, since it requires $\mathbf{P}(t_i^{+})$ to be known before $\mathbf{K}(t_i)$ can be obtained! However, this equivalent expression will be of use for the continuous-measurement case and in fact is the computational form of the Kalman gain for that case.

5.7 INVERSE COVARIANCE FORM

The previous section mentioned the idea of expressing the optimal estimation algorithm in terms of the inverse of the covariance matrix, instead of the covariance itself. Though an algebraically equivalent result, this form will possess some unique characteristics, as allowing a startup procedure for the case of $\mathbf{P}_0^{-1}$ being singular. This form will be directly exploited in the optimal smoothers of Chapter 8 (Volume 2). Moreover, the inverse covariance matrix is directly related to the Fisher information matrix, allowing an interpretation of filter performance in terms of information theoretic concepts. Finally, the relationship of the optimal estimator to the classical Gauss–Markov theorem for a special class of problems becomes apparent from this form.

The usual recursion relations for the covariance matrix in the estimation algorithm can be written as

$$\mathbf{P}^{-1}(t_i{}^+) = \mathbf{P}^{-1}(t_i{}^-) + \mathbf{H}^{\mathrm{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i) \tag{5-85}$$

$$\mathbf{P}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i{}^+)\mathbf{\Phi}^{\mathrm{T}}(t_{i+1}, t_i) + \mathbf{G_d}(t_i)\mathbf{Q_d}(t_i)\mathbf{G_d}^{\mathrm{T}}(t_i) \tag{5-86}$$

Applying the matrix inversion lemma to (5-85) yields the familiar Kalman filter equations. Instead, apply the lemma, one form of which states that for $\mathbf{X}$ and $\mathbf{Y}$ both n-by-k matrices,

$$(\mathbf{A} + \mathbf{X}^{\mathrm{T}}\mathbf{Y})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I} + \mathbf{Y}\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{Y}\mathbf{A}^{-1} \tag{5-87}$$

to (5-86) by identifying

$$\mathbf{A} = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i{}^+)\mathbf{\Phi}^{\mathrm{T}}(t_{i+1}, t_i); \qquad \mathbf{X}^{\mathrm{T}} = \mathbf{G_d}(t_i)\mathbf{Q_d}(t_i); \qquad \mathbf{Y} = \mathbf{G_d}^{\mathrm{T}}(t_i)$$

to yield, for $\mathbf{Q_d}(t_i)$ nonsingular (not very restrictive),

$$\mathbf{P}^{-1}(t_{i+1}^-) = \mathbf{M}(t_{i+1}) - \mathbf{M}(t_{i+1})\mathbf{G_d}(t_i)[\mathbf{G_d}^{\mathrm{T}}(t_i)\mathbf{M}(t_{i+1})\mathbf{G_d}(t_i)$$
$$+ \mathbf{Q_d}^{-1}(t_i)]^{-1}\mathbf{G_d}^{\mathrm{T}}(t_i)\mathbf{M}(t_{i+1}) \tag{5-88}$$

where

$$\mathbf{M}(t_{i+1}) = \mathbf{\Phi}^{\mathrm{T}}(t_i, t_{i+1})\mathbf{P}^{-1}(t_i{}^+)\mathbf{\Phi}(t_i, t_{i+1}) \tag{5-89}$$

In (5-89), note the order of time indices in the state transition matrices: appropriate for backward time propagation of the system relations (forward propagation of adjoint relations). If $\mathbf{Q_d}(t_i) \equiv \mathbf{0}$, then (5-88) is not applicable, and $\mathbf{P}^{-1}(t_{i+1}) = \mathbf{M}(t_{i+1})$. The inverse covariance time propagation equation (5-88) is analogous in form to the covariance update equation (5-79). The analog of the Joseph form (5-82) has also been derived, with superior numerical characteristics similar to that of (5-82). Define the gain matrix $\mathcal{X}(t_i)$ as

$$\mathcal{X}(t_i) = \mathbf{M}(t_{i+1})\mathbf{G_d}(t_i)[\mathbf{G_d}^{\mathrm{T}}(t_i)\mathbf{M}(t_{i+1})\mathbf{G_d}(t_i) + \mathbf{Q_d}^{-1}(t_i)]^{-1} \tag{5-90}$$

In terms of this gain, (5-88) becomes

$$\mathbf{P}^{-1}(t_{i+1}^-) = \mathbf{M}(t_{i+1}) - \mathcal{X}(t_i)\mathbf{G_d}^{\mathrm{T}}(t_i)\mathbf{M}(t_{i+1}) \tag{5-91}$$

or, in the analog of the Joseph form [49],

$$\mathbf{P}^{-1}(t_{i+1}^-) = [\mathbf{I} - \mathcal{X}(t_i)\mathbf{G_d}^{\mathrm{T}}(t_i)]\mathbf{M}(t_{i+1})[\mathbf{I} - \mathcal{X}(t_i)\mathbf{G_d}^{\mathrm{T}}(t_i)]^{\mathrm{T}} + \mathcal{X}(t_i)\mathbf{Q_d}^{-1}(t_i)\mathcal{X}^{\mathrm{T}}(t_i) \tag{5-92}$$

In certain circumstances, the a priori statistical information about the state may not be complete: there is no information about the state initial conditions in some or all directions of state space. This can be modeled as the limiting case of certain eigenvalues of $\mathbf{P}_0$ going to infinity, or those of $\mathbf{P}_0^{-1}$ going to zero. Because it remains finite, the inverse covariance would be more desirable to employ.

If $\mathbf{P}^{-1}(t_0) = \mathbf{P}_0^{-1}$ is singular, then until $\mathbf{P}^{-1}(t_i)$ attains full rank, a unique estimate of the full state cannot be made. To allow a viable startup procedure, the state estimates $\hat{\mathbf{x}}(t_i{}^-)$ and $\hat{\mathbf{x}}(t_i{}^+)$ are replaced by

$$\hat{\mathbf{y}}(t_i{}^-) \triangleq \mathbf{P}^{-1}(t_i{}^-)\hat{\mathbf{x}}(t_i{}^-) \tag{5-93a}$$

$$\hat{\mathbf{y}}(t_i{}^+) \triangleq \mathbf{P}^{-1}(t_i{}^+)\hat{\mathbf{x}}(t_i{}^+) \tag{5-93b}$$

The recursions for $\hat{\mathbf{y}}$ are then

$$\hat{\mathbf{y}}(t_i{}^+) = \hat{\mathbf{y}}(t_i{}^-) + \mathbf{H}^{\mathrm{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{z}_i \tag{5-94}$$

$$\hat{\mathbf{y}}(t_{i+1}^-) = [\mathbf{I} - \mathscr{X}(t_i)\mathbf{G}_{\mathrm{d}}{}^{\mathrm{T}}(t_i)]\mathbf{\Phi}^{\mathrm{T}}(t_i, t_{i+1})[\hat{\mathbf{y}}(t_i{}^+) + \mathbf{P}^{-1}(t_i{}^+)\mathbf{\Phi}(t_i, t_{i+1})\mathbf{B}_{\mathrm{d}}(t_i)\mathbf{u}(t_i)] \tag{5-95}$$

starting from the initial condition

$$\hat{\mathbf{y}}(t_0) = \mathbf{P}_0^{-1}\hat{\mathbf{x}}_0 \tag{5-96}$$

Once $\mathbf{P}^{-1}(t_i{}^+)$ becomes nonsingular, then its inverse can be computed to obtain $\mathbf{P}(t_i{}^+)$, and the optimal state estimate can be expressed as

$$\hat{\mathbf{x}}(t_i{}^+) = \mathbf{P}(t_i{}^+)\hat{\mathbf{y}}(t_i{}^+) \tag{5-97}$$

From that time forward, it is possible to revert to the more familiar covariance form or to continue in the inverse covariance form. The latter is more convenient in certain situations, as in smoothing.

Thus, the inverse covariance form of the optimal estimator has been described. Measurement updating is accomplished through (5-85) and (5-94), and time propagation by (5-89), (5-90), (5-91) or (5-92), and (5-95). Initial conditions are $\mathbf{P}^{-1}(t_0) = \mathbf{P}_0^{-1}$ and $\hat{\mathbf{y}}(t_0) = \mathbf{P}_0^{-1}\hat{\mathbf{x}}_0$. These relations are valid for $\mathbf{Q}_{\mathrm{d}}(t_i)$ positive definite; unless there is no driving noise, a positive definite s-by-s $\mathbf{Q}_{\mathrm{d}}(t_i)$ can always be generated for an n-by-n positive semidefinite $[\mathbf{G}_{\mathrm{d}}(t_i)\mathbf{Q}_{\mathrm{d}}(t_i)\mathbf{G}_{\mathrm{d}}{}^{\mathrm{T}}(t_i)]$ form of rank s. If there is no driving noise $[\mathbf{Q}_{\mathrm{d}}(t_i) = \mathbf{0}$ for all $t_i]$, then the time propagation relations become

$$\mathbf{P}^{-1}(t_{i+1}^-) = \mathbf{M}(t_{i+1}) \tag{5-98}$$

$$\hat{\mathbf{y}}(t_{i+1}^-) = \mathbf{\Phi}^{\mathrm{T}}(t_i, t_{i+1})[\hat{\mathbf{y}}(t_i{}^+) + \mathbf{P}^{-1}(t_i{}^+)\mathbf{\Phi}(t_i, t_{i+1})\mathbf{B}_{\mathrm{d}}(t_i)\mathbf{u}(t_i)] \tag{5-99}$$

A concept related to the inverse covariance is the *Fisher information matrix* which is a measure of the certainty of the state estimate due to measurement data *alone*; i.e., the a priori information of $\mathbf{x}(t_0)$ being modeled as Gaussian with mean $\hat{\mathbf{x}}_0$ and covariance $\mathbf{P}_0$ is disregarded. The information matrix $\mathscr{I}(t_i, t_1)$ is given by

$$\mathscr{I}(t_i, t_1) = \sum_{j=1}^{i} \mathbf{\Phi}^{\mathrm{T}}(t_j, t_i)\mathbf{H}^{\mathrm{T}}(t_j)\mathbf{R}^{-1}(t_j)\mathbf{H}(t_j)\mathbf{\Phi}(t_j, t_i) \tag{5-100}$$

where, as noted previously, $\boldsymbol{\Phi}(t_j, t_i)$ for $j < i$ is the transition matrix for propagating the system state backward in time. To relate this concept directly to the previous algorithm, ignore the dynamics driving noise: assume that there is no $\mathbf{w}_d(t_i)$ sequence, or equivalently, that $\mathbf{Q}_d(t_i) = \mathbf{0}$ for all t_i. Under this assumption, (5-85), (5-89), and (5-98) yield

$$\mathscr{I}(t_i, t_1) = \mathbf{P}^{-1}(t_i{}^+) - \boldsymbol{\Phi}^T(t_0, t_i)\mathbf{P}_0^{-1}\boldsymbol{\Phi}(t_0, t_i) \qquad (5\text{-}101)$$

If there were no a priori information about the state, or formally if $\mathbf{P}_0^{-1} = \mathbf{0}$, then the information matrix is the inverse of the corresponding estimator error covariance. The larger the eigenvalues of $\mathscr{I}(t_i, t_1)$, the smaller the eigenvalues of $\mathbf{P}(t_i{}^+)$, and the more precise our estimate is. If any eigenvalues of $\mathscr{I}(t_i, t_1)$ are zero, there are directions in state space along which our measurements give us no information. Not surprisingly, this information matrix is directly related to the observability matrix studied in Chapter 2 [see Eq. (2-74)].

Expressing the definition of the information matrix, (5-100), for times t_i and t_{i-1}, and equating like terms, yields the following recursion:

$$\mathscr{I}(t_i, t_1) = \boldsymbol{\Phi}^T(t_{i-1}, t_i)\mathscr{I}(t_{i-1}, t_1)\boldsymbol{\Phi}(t_{i-1}, t_i) + \mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i) \quad (5\text{-}102)$$

From this relation, it can be seen that the "information" contained in a single measurement at time t_i is $[\mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)]$: the term added to $\mathbf{P}^{-1}(t_i{}^-)$ to generate $\mathbf{P}^{-1}(t_i{}^+)$.

EXAMPLE 5.8   Reconsider the estimator for the gyro on test, Example 5.2, in inverse variance form. Initial conditions are

$$P^{-1}(t_0) = P_0^{-1} = 1 \quad \text{hr}^2/\text{deg}^2$$
$$\hat{y}(t_0) = P_0^{-1}\hat{x}_0 = 0$$

Equations (5-85) and (5-94) yield the measurement updates as

$$P^{-1}(t_i{}^+) = P^{-1}(t_i{}^-) + H^T R^{-1} H = P^{-1}(t_i{}^-) + 2$$
$$\hat{y}(t_i{}^+) = \hat{y}(t_i{}^-) + H^T R^{-1} z_i = \hat{y}(t_i{}^-) + 2z_i$$

The time propagations are

$$M(t_{i+1}) = (\Phi^{-1})^T P^{-1}(t_i{}^+)\Phi^{-1} = (0.78)^{-2} P^{-1}(t_i{}^+) = 1.64 P^{-1}(t_i{}^+)$$

$$\mathscr{X}(t_i) = \frac{M(t_{i+1})G_d}{G_d{}^T M(t_{i+1})G_d + Q_d{}^{-1}} = \frac{M(t_{i+1})}{M(t_{i+1}) + (0.39)^{-1}} = \frac{M(t_{i+1})}{M(t_{i+1}) + 2.56}$$

$$P^{-1}(t_{i+1}^-) = M(t_{i+1}) - \mathscr{X}(t_i)G_d{}^T M(t_{i+1})$$

$$\hat{y}(t_{i+1}^-) = [1 - \mathscr{X}(t_i)G_d{}^T](\Phi^{-1})^T \hat{y}(t_i{}^+) = 1.28[1 - \mathscr{X}(t_i)]\hat{y}(t_i{}^+)$$

The time history of the inverse variance (and gains) can be precomputed, and is displayed in Fig. 5.10. These results are directly comparable to Fig. 5.6, the plot of the error variance for the same problem. Note that the "information" added at each measurement time is $H^T R^{-1} H = 2 \text{ hr}^2/\text{deg}^2$.   ∎
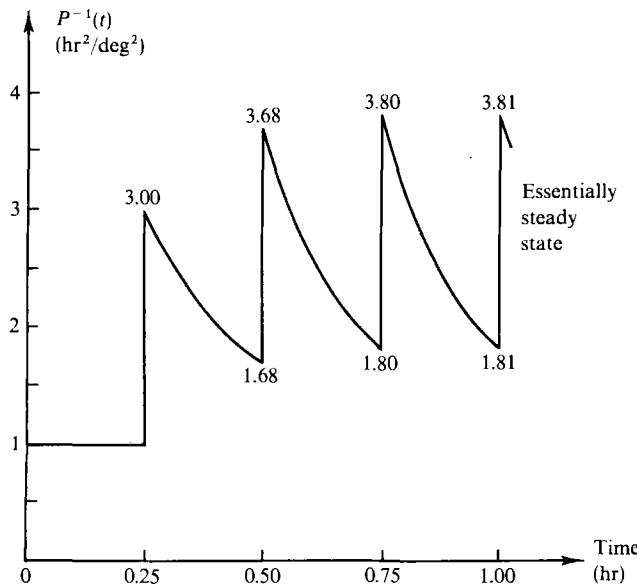
FIG. 5.10   Inverse variance time history for Example 5.8.

5.8   STABILITY

This section specifies the rather nonrestrictive conditions under which the filter algorithm is stable [22, 53]. Issues involved in the stability of stochastically driven systems are not completely resolved, but Lyapunov stability theory has been applied to the homogeneous portion of the filter to establish *zero-input stability* criteria. In fact, conditions will be established under which the filter is a uniformly, asymptotically stable linear system, which then implies *bounded input–bounded output* (BIBO) *stability*: for bounded inputs into the filter, the output (state estimate) is bounded. We note that zero-input and BIBO stability are significantly more distinct issues for nonlinear systems.

The state equations (5-36) and (5-39) of the Kalman filter algorithm can be rewritten as

$$\hat{\mathbf{x}}(t_i^+) = [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)\mathbf{z}_i$$
$$= [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+)$$
$$+ [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)] \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau + \mathbf{K}(t_i)\mathbf{z}_i \quad (5\text{-}103)$$

We want to consider the stability of the homogeneous part of the filter,

$$\hat{\mathbf{x}}_{\mathbf{h}}(t_i^+) = [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}_{\mathbf{h}}(t_{i-1}^+) \qquad (5\text{-}104)$$

a linear discrete system model with state transition matrix equal to

$$[\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\mathbf{\Phi}(t_i, t_{i-1}).$$

To specify the sufficient conditions for stability succinctly, we must introduce some system theory concepts and terminology. We assume a system model as described in Section 5.2. Such a system representation is said to be *stochastically controllable* if there exist positive numbers α and β, $0 < \alpha < \beta < \infty$, and a time interval Δt such that, for all $t \geq t_0 + \Delta t$,

$$\alpha\mathbf{I} \leq \int_{t-\Delta t}^{t} \mathbf{\Phi}(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\mathbf{\Phi}^T(t, \tau)\, d\tau \leq \beta\mathbf{I} \qquad (5\text{-}105)$$

where $\mathbf{M}_1 \geq \mathbf{M}_2$ means $(\mathbf{M}_1 - \mathbf{M}_2) \geq \mathbf{0}$, i.e., $(\mathbf{M}_1 - \mathbf{M}_2)$ is positive semidefinite. This implies that the system is completely controllable with respect to the points of entry of the dynamic driving noise (see Section 2.5), which further implies that the driving noise affects all of the states. However, (5-105) is a stricter requirement than complete controllability: not only must the integral be positive definite, but must be bounded both above and below.

Analogously, the discrete-time system representation (5-49), which may have arisen as an equivalent discrete-time system model, is stochastically controllable if there exist α and β, $0 < \alpha < \beta < \infty$, and a positive integer N such that, for all $i \geq N$,

$$\alpha\mathbf{I} \leq \sum_{j=i-N+1}^{i} \mathbf{\Phi}(t_i, t_j)\mathbf{G}_d(t_{j-1})\mathbf{Q}_d(t_{j-1})\mathbf{G}_d^T(t_{j-1})\mathbf{\Phi}^T(t_i, t_j) \leq \beta\mathbf{I} \quad (5\text{-}106)$$

As in (5-105), this is stricter than, and implies, complete controllability with respect to the points of entry of the dynamic driving noise $\mathbf{w}_d(\cdot, \cdot)$.

The sampled-data system representation of Section 5.2 is said to be *stochastically observable* if there exist positive numbers α and β, $0 < \alpha < \beta < \infty$, and a positive integer N such that, for all $i \geq N$,

$$\alpha\mathbf{I} \leq \sum_{j=i-N+1}^{i} \mathbf{\Phi}^T(t_j, t_i)\mathbf{H}^T(t_j)\mathbf{R}^{-1}(t_j)\mathbf{H}(t_j)\mathbf{\Phi}(t_j, t_i) \leq \beta\mathbf{I} \qquad (5\text{-}107)$$

Due to the requirement of being bounded both above and below, this is a stronger condition than, and implies, complete observability with respect to the points of exit of the measurements from the system model. Thus, it also implies that the effects of changes of any states can be observed in the outputs. Note that the summation term that appears in (5-107) is in fact the information matrix $\mathscr{I}(t_i, t_{i-N+1})$.

In Section 5.10, optimal estimation for the case of continuously available measurements will be discussed. The measurement model will be

$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}(t) \qquad (5\text{-}108)$$

with $\mathbf{v}(\cdot, \cdot)$ a zero-mean white Gaussian noise with $E\{\mathbf{v}(t)\mathbf{v}^{\mathrm{T}}(t + \tau)\} = \mathbf{R}_{\mathrm{c}}(t)\delta(\tau)$. Such a system representation is similarly said to be stochastically observable if there exist positive numbers α and β, $0 < \alpha < \beta < \infty$, and a time interval Δt such that, for all $t \geq t_0 + \Delta t$,

$$\alpha\mathbf{I} \leq \int_{t-\Delta t}^{t} \mathbf{\Phi}^{\mathrm{T}}(\tau, t)\mathbf{H}^{\mathrm{T}}(\tau)\mathbf{R}_{\mathrm{c}}^{-1}(\tau)\mathbf{H}(\tau)\mathbf{\Phi}(\tau, t)\, d\tau \leq \beta\mathbf{I} \qquad (5\text{-}109)$$

The integral in this expression is the information matrix, $\mathscr{I}(t, t - \Delta t)$, appropriate to this continuous-time model.

Note that the condition of stochastic controllability is not met if $\mathbf{Q}(\tau) \equiv \mathbf{0}$ over the entire interval in (5-105) or $\mathbf{Q}_{\mathrm{d}}(t_{j-1}) \equiv \mathbf{0}$ over the summation range in (5-106): the case of no dynamic driving noise. Neither is it met if these strengths have infinite eigenvalues over a finite length of time. Similarly, stochastic observability is violated if $\mathbf{R}^{-1}(t_j)$ or $\mathbf{R}_{\mathrm{c}}^{-1}(\tau)$ are zero over the entire time of interest (infinite noise corruption) or if they are infinite over any finite time (the case of perfect measurements).

If the system model upon which the Kalman filter is based is stochastically observable and stochastically controllable, then the filter is uniformly asymptotically globally stable. This means that if we consider the homogeneous equation (5-104), then

$$\lim_{i \to \infty} \|\hat{\mathbf{x}}_{\mathrm{h}}(t_i^+)\| = 0 \qquad (5\text{-}110)$$

i.e., in the limit as the number of data samples grows without bound, the norm ("length" or magnitude) of $\hat{\mathbf{x}}_{\mathrm{h}}(t_i^+)$ goes to zero (asymptotic), no matter what the initial conditions (global), and the rate of convergence is not a function of absolute time (uniform). Mathematically, the system model is uniformly asymptotically stable if there exist positive constants α and β such that, for *all* $t_i \geq t_0$, the state transition matrix to transition $\hat{\mathbf{x}}_{\mathrm{h}}(t_0)$ to time t_i has a norm bounded above by $\alpha \exp[-\beta(t_i - t_0)]$. The proof of this claim through explicit generation of an appropriate Lyapunov function is omitted, but can be found in the work of Kalman [36], Deyst and Price [22], Sorenson [72], Jazwinski [31], and McGarty [53] (the last reference correcting errors made in previous derivations).

It is important that the sufficient conditions for filter stability do not include stability of the original system model itself. The system model (and the actual system itself) can be unstable, and the filter equations may simultaneously be stable. Even if the system states are in fact growing without bound, as for example in the onset of nuclear reactor runaway, the errors committed by the filter in estimating those states will remain bounded. Thus, if the "true" state time history were as in Fig. 5.11, the state estimate could track this behavior as in plot (a) rather than exhibit an error growing unbounded so as to indicate no system instability, as in plot (b). This is a very desirable filter characteristic.
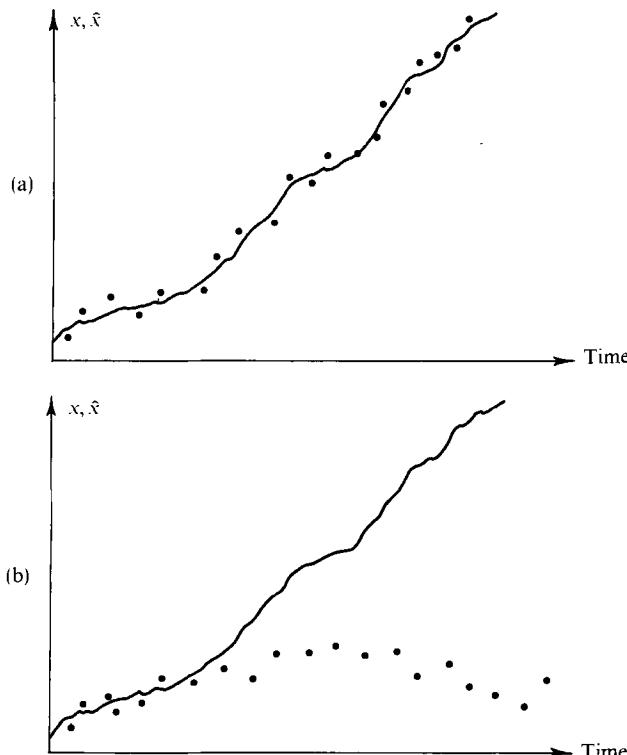
FIG. 5.11   Estimation performance for unstable systems. (a) Unstable system and stable filter. (b) Unstable system and unstable filter. The solid lines indicate "true state" and the dots indicate state estimates.

As might be suggested by the filter derivation of Section 5.3, if the system model is stochastically controllable, then $\mathbf{P}(t_i^+)$ is positive definite for all $i \geq N$. If it is both stochastically controllable and stochastically observable, $\mathbf{P}(t_i^+)$ is also uniformly bounded from above for all $i \geq N$. Furthermore, under these conditions, if $\mathbf{P}_1(t_i^+)$ and $\mathbf{P}_2(t_i^+)$ are two solutions to the filter recursions for different initial conditions $\mathbf{P}_{10} \geq 0$ and $\mathbf{P}_{20} \geq 0$, respectively, then $[\mathbf{P}_1(t_i^+) - \mathbf{P}_2(t_i^+)]$ converges uniformly asymptotically globally to $\mathbf{0}$. (For proofs, see Jazwinski [31].) This last claim indicates that as more measurement information is incorporated, the effect of $\mathbf{P}_0$ (which is often subject to uncertainty itself) is "forgotten." Since the same would be true of the $[\mathbf{P}_1(t_i^+) - \mathbf{P}_2(t_i^+)]$ sequence when started from some time later than t_0, this also indicates that the effect of numerical errors in computing $\mathbf{P}(t_i^+)$ are similarly "forgotten." However, numerical errors associated with finite computer wordlength warrant particular attention, in that the stability claims just made are based upon an assumption of unbounded wordlength. Chapter 7 will discuss this problem and its solution in greater detail.

5.9 CORRELATION OF DYNAMIC DRIVING NOISE
AND MEASUREMENT NOISE

Assume that a sampled-data problem is adequately described by the discrete-time (possibly equivalent discrete-time) system model

$$\mathbf{x}(t_i) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{x}(t_{i-1}) + \mathbf{B_d}(t_{i-1})\mathbf{u}(t_{i-1}) + \mathbf{G_d}(t_{i-1})\mathbf{w_d}(t_{i-1}) \quad (5\text{-}49)$$

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{5-4}$$

with the usual Gaussian description of the initial conditions and zero-mean white noise processes $\mathbf{w_d}(\cdot,\cdot)$ and $\mathbf{v}(\cdot,\cdot)$:

$$E\{\mathbf{x}(t_0)\} = \hat{\mathbf{x}}_0, \qquad E\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}_0][\mathbf{x}(t_0) - \hat{\mathbf{x}}_0]^\mathrm{T}\} = \mathbf{P}_0 \tag{5-3}$$

$$E\{\mathbf{w_d}(t_i)\mathbf{w_d}^\mathrm{T}(t_j)\} = \begin{cases} \mathbf{Q_d}(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{5-50}$$

$$E\{\mathbf{v}(t_i)\mathbf{v}^\mathrm{T}(t_j)\} = \begin{cases} \mathbf{R}(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{5-5}$$

The Kalman filter for this problem formulation was previously investigated under the assumption that the dynamic driving noise $\mathbf{w_d}(\cdot,\cdot)$ and measurement noise $\mathbf{v}(\cdot,\cdot)$ were uncorrelated; the algorithm is specified by (5-38)–(5-42) and (5-51)–(5-52).

Now we want to consider an extension of these results, allowing correlation between the two noise processes (assumed jointly Gaussian), the need for which was motivated in Section 5.2. Let this correlation be described by

$$E\{\mathbf{w_d}(t_i)\mathbf{v}^\mathrm{T}(t_j)\} = \begin{cases} \mathbf{C}(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{5-111}$$

A generalized derivation [36] yields the optimal estimation algorithm with the same initial conditions and measurement update relations, i.e., leaving (5-38)–(5-42) unaltered, but with time propagation equations modified from (5-51) and (5-52) to (note the time index has been changed for convenience):

$$\hat{\mathbf{x}}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\hat{\mathbf{x}}(t_i^+) + \mathbf{B_d}(t_i)\mathbf{u}(t_i)$$
$$\qquad + \mathbf{G_d}(t_i)\mathbf{C}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^\mathrm{T}(t_i) + \mathbf{R}(t_i)]^{-1}[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \quad (5\text{-}112)$$

$$\mathbf{P}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i^+)\mathbf{\Phi}^\mathrm{T}(t_{i+1}, t_i) + \mathbf{G_d}(t_i)\mathbf{Q_d}(t_i)\mathbf{G_d}^\mathrm{T}(t_i)$$
$$\qquad - \mathbf{G_d}(t_i)\mathbf{C}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^\mathrm{T}(t_i) + \mathbf{R}(t_i)]^{-1}\mathbf{C}^\mathrm{T}(t_i)\mathbf{G_d}^\mathrm{T}(t_i)$$
$$\qquad - \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{K}(t_i)\mathbf{C}^\mathrm{T}(t_i)\mathbf{G_d}^\mathrm{T}(t_i) - \mathbf{G_d}(t_i)\mathbf{C}(t_i)\mathbf{K}^\mathrm{T}(t_i)\mathbf{\Phi}^\mathrm{T}(t_{i+1}, t_i) \quad (5\text{-}113)$$

Note that if t_1 is assumed to be the first measurement time, (5-51) and (5-52) are used for the first sample period.

EXAMPLE 5.9   Return to the gyro on test example, but now let

$$E\{\mathbf{w_d}(t_i)\mathbf{v}(t_i)\} = C = 0.2$$

Then the time propagation relations become (note $G_d = 1$)

$$\hat{x}(t_{i+1}^-) = 0.78\hat{x}(t_i^+) + 0.2[P(t_i^-) + 0.5]^{-1}[z_i - \hat{x}(t_i^-)]$$

$$P(t_{i+1}^-) = (0.78)^2 P(t_i^+) + 0.39 - (0.2)^2[P(t_i^-) + 0.5]^{-1} - 2[0.78]K(t_i)[0.2]$$

or, since $K(t_i) = P(t_i^+)H(t_i)R(t_i)^{-1} = 2P(t_i^+)$,

$$P(t_{i+1}^-) = 0.78(0.78 - 0.80)P(t_i^+) + 0.39 - 0.04[P(t_i^-) + 0.5]^{-1}$$

$$= -0.016P(t_i^+) + 0.39 - 0.04[P(t_i^-) + 0.5]^{-1}$$

Calculating the error variance and gain time history yields

Time	0	0.25	0.50	0.75	1.00
$P(t_i^-)$	–	1.00	0.35	0.34	0.34
$P(t_i^+)$	1	0.33	0.21	0.20	0.20
$K(t_i)$	–	0.67	0.42	0.40	0.40

This can be compared to Fig. 5.6 for the case of no correlation between $\mathbf{w_d}(\cdot,\cdot)$ and $\mathbf{v}(\cdot,\cdot)$. The decrease in steady state values from $P(t_i^-) = 0.55$ to 0.34 and $P(t_i^+)$ from 0.26 to 0.20 is due to the exploitation of the correlation between the dynamic noise and the noise that corrupts the observable outputs: the z_i realizations reveal more about the noise process $\mathbf{w_d}(\cdot,\cdot)$.   ∎

If (5-49) is an equivalent discrete-time system model, the previous problem formulation corresponds to correlation between the noise corrupting the measurement at time t_i and the dynamic driving noise over the ensuing sample period. It is also useful to consider a model involving correlation between the dynamic noise over a sample period and the noise corrupting the measurement at the end of that interval. Thus, consider the same problem formulation, but replacing (5-11) with

$$E\{\mathbf{w_d}(t_{i-1})\mathbf{v}^T(t_j)\} = \begin{cases} \mathbf{C}(t_j) & t_j = t_i \\ 0 & t_j \neq t_i \end{cases} \qquad (5\text{-}114)$$

For this formulation, the generalized filter algorithm entails the original initial conditions and time propagation equations, i.e., (5-41), (5-42), (5-51), and (5-52). However, the measurement update relations (5-38)–(5-40) are replaced by (note the similarity to the result of Problem 3.19)

$$\mathbf{K_c}(t_i) = [\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{G_d}(t_{i-1})\mathbf{C}(t_i)][\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i)$$
$$+ \mathbf{R}(t_i) + \mathbf{H}(t_i)\mathbf{G_d}(t_{i-1})\mathbf{C}(t_i) + \mathbf{C}^T(t_i)\mathbf{G_d}^T(t_{i-1})\mathbf{H}^T(t_i)]^{-1} \quad (5\text{-}115)$$

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K_c}(t_i)[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \qquad (5\text{-}116)$$

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{K_c}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-) + \mathbf{C}^T(t_i)\mathbf{G_d}^T(t_{i-1})] \qquad (5\text{-}117)$$

EXAMPLE 5.10   Again consider the gyro on test example with correlation between $w_d(\cdot,\cdot)$ and $v(\cdot,\cdot)$, but now in the form of

$$E\{w_d(t_{i-1})v(t_i)\} = C = 0.2$$

Then the measurement update equations become (with $G_d = 1$):

$$K_c(t_i) = [P(t_i^-) + 0.2][P(t_i^-) + 0.5 + 0.2 + 0.2]^{-1} = [P(t_i^-) + 0.2]/[P(t_i^-) + 0.9]$$
$$\hat{x}(t_i^+) = \hat{x}(t_i^-) + K_c(t_i)[z_i - x(t_i^-)]$$
$$P(t_i^+) = P(t_i^-) - K_c(t_i)[P(t_i^-) + 0.2]$$

The error variance and gain time histories can be computed as

Time	0	0.25	0.50	0.75	1.00
$P(t_i^-)$	–	1.00	0.54	0.49	0.48
$P(t_i^+)$	1	0.24	0.16	0.15	0.15
$K_c(t_i)$	–	0.63	0.51	0.50	0.49

Here there is no correlation between the $v(t_i)$ corrupting the current measurement and $w_d(t_i)$ driving the dynamics during the next sample period, so there is a greater spreading of errors over the next interval than in Example 5.9: in steady state, $[P(t_{i+1}^-) - P(t_i^+)]$ here is $(0.48 - 0.15) = 0.33$ versus $(0.34 - 0.20) = 0.14$. With less confidence in the dynamics model here, the filter gain is higher (0.49 versus 0.40 in steady state) to weight the measurement data more heavily. By being correlated with $w_d(t_{i-1})$, $v(t_i)$ is also correlated with $x(t_i)$, so the measurement $z(t_i)$ is more strongly correlated with $x(t_i)$, and thus the lower $P(t_i^+)$ value here (0.15 versus 0.20).   ∎

These examples reveal the improved estimation precision due to exploiting the noise correlation: observing a particular realization of $v(\cdot,\cdot)$ as the corruption on the measurements yields probabilistic information about the realizations of $w_d(\cdot,\cdot)$ that have driven the system dynamics. As expected, either of these formulations reduce to the original Kalman filter algorithm if $C(t_i) \equiv 0$ for all time. Moreover, if $v(t_i)$ is correlated with both $w_d(t_{i-1})$ and $w_d(t_i)$, as

$$E\left\{\begin{bmatrix} w_d(t_{i-1}) \\ w_d(t_i) \\ v(t_i) \end{bmatrix}[w_d^T(t_{i-1})w_d^T(t_i)v^T(t_i)]\right\} = \begin{bmatrix} Q_d(t_{i-1}) & 0 & C_1(t_i) \\ 0 & Q_d(t_i) & C_2(t_i) \\ C_1^T(t_i) & C_2^T(t_i) & R(t_i) \end{bmatrix}$$

the above results can be combined. In practical applications, a tradeoff analysis would be conducted to determine whether the performance improvement afforded by this algorithm warrants the increased computer burden beyond that of the more conventional Kalman filter formulation.

5.10   TIME-CORRELATED MEASUREMENT NOISE; PERFECT MEASUREMENTS

In the optimal estimator derivation of Section 5.3, the covariance of the measurement corruption noise, $R(t_i)$, was assumed positive definite because $R^{-1}(t_i)$ was required in certain steps of that derivation. However, $R^{-1}(t_i)$ does not ap-

pear in the final recursions, so positive definiteness may not in fact be necessary. In fact, the algorithm will operate as long as $[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}(t_i) + \mathbf{R}(t_i)]$ is invertible. The particular case of $\mathbf{R}(t_i) \equiv \mathbf{0}$ for all time is of interest because of its applicability to problems in which measurements are corrupted only by time-correlated noise as well as to formulations involving "perfect" measurements. For these cases in which $\mathbf{R}(t_i)$ is singular, stochastic observability may well be violated, but recall that this is part of a sufficient rather than necessary condition for filter stability. Numerical problems will be accentuated, however (see Chapter 7).

First let us demonstrate that time-correlated measurement noise does lead to a filter formulation in which $\mathbf{R}(t_i)$ is zero for all time. Let the system model be described (in white noise notation) as in Fig. 5.12. The system dynamics model is

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \qquad (5\text{-}118)$$

with $\mathbf{w}(\cdot,\cdot)$ a zero-mean white Gaussian noise of strength $\mathbf{Q}(t)$ for all $t \in T$. The available discrete-time measurements are modeled as

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{n}_f(t_i) \qquad (5\text{-}119)$$

where $\mathbf{n}_f(\cdot,\cdot)$ is a zero-mean time-correlated Gaussian process: the output of a time-correlated noise generator, i.e., a shaping filter. This shaping filter is described by (the subscript f denotes filter)

$$\dot{\mathbf{x}}_f(t) = \mathbf{F}_f(t)\mathbf{x}_f(t) + \mathbf{G}_f(t)\mathbf{w}_f(t) \qquad (5\text{-}120)$$

$$\mathbf{n}_f(t) = \mathbf{H}_f(t)\mathbf{x}_f(t) \qquad (5\text{-}121)$$

where $\mathbf{w}_f(\cdot,\cdot)$ is a zero-mean white Gaussian noise of strength $\mathbf{Q}_f(t)$ for all $t \in T$.
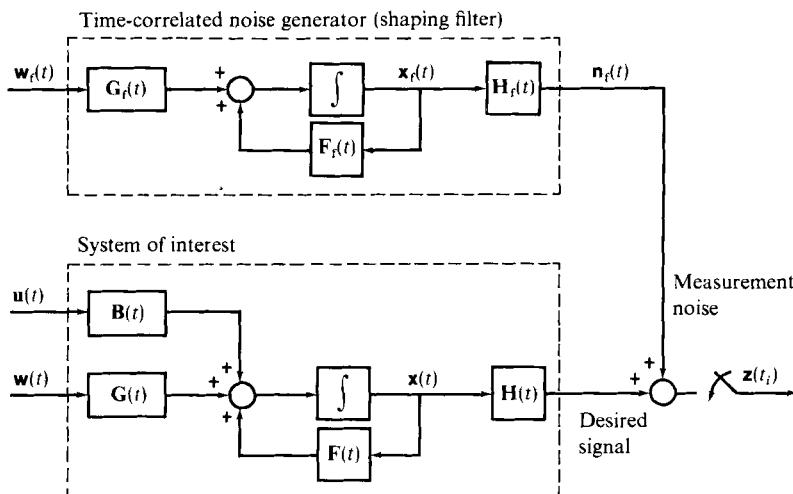
FIG. 5.12   Time-correlated measurement noise model.

The overall system model can be expressed in terms of the augmented state $\mathbf{x}_a(\cdot, \cdot)$ as described in Section 4.11,

$$\mathbf{x}_a(\cdot, \cdot) = \left[\begin{array}{c} \mathbf{x}(\cdot, \cdot) \\ \hline \mathbf{x}_f(\cdot, \cdot) \end{array} \right] \tag{5-122}$$

as

$$\left[\begin{array}{c} \dot{\mathbf{x}}(t) \\ \hline \dot{\mathbf{x}}_f(t) \end{array} \right] = \left[\begin{array}{c|c} \mathbf{F}(t) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{F}_f(t) \end{array} \right] \left[\begin{array}{c} \mathbf{x}(t) \\ \hline \mathbf{x}_f(t) \end{array} \right] + \left[\begin{array}{c} \mathbf{B}(t) \\ \hline \mathbf{0} \end{array} \right] \mathbf{u}(t) + \left[\begin{array}{c|c} \mathbf{G}(t) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{G}_f(t) \end{array} \right] \left[\begin{array}{c} \mathbf{w}(t) \\ \hline \mathbf{w}_f(t) \end{array} \right]$$

$$\tag{5-123}$$

$$\mathbf{Q}_a(t) = \left[\begin{array}{c|c} \mathbf{Q}(t) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Q}_f(t) \end{array} \right] \tag{5-124}$$

$$\mathbf{z}(t_i) = [\mathbf{H}(t_i) \mid \mathbf{H}_f(t_i)] \left[\begin{array}{c} \mathbf{x}(t_i) \\ \hline \mathbf{x}_f(t_i) \end{array} \right] \tag{5-125}$$

This is in the form of a linear system driven only by white Gaussian noise $\mathbf{w}_a(\cdot, \cdot)$, but from which is available perfect measurements of certain linear combinations of states at discrete times.

Now let us consider the problem involving perfect measurements in general. The optimal estimator will be generated by a limiting process on previous results, and the difference in the characteristics of the resulting algorithm will be described. Let the available measurements be modeled as

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i)$$

with the zero-mean white Gaussian noise $\mathbf{v}(\cdot, \cdot)$ being of strength $\mathbf{R}(t_i) = \varepsilon \mathbf{I}$, $\varepsilon > 0$ (so that $\mathbf{R}(t_i)$ is positive definite). The filter update equations are

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \varepsilon \mathbf{I}]^{-1}[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$$
$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \varepsilon \mathbf{I}]^{-1}\mathbf{H}(t_i)\mathbf{P}(t_i^-)$$

Now observe the result of letting $\varepsilon \to 0$, i.e., the limit of perfect measurements. Although ε must be nonzero to be assured of the existence of density functions used in the original derivation, characteristic functions can be used to maintain validity of this operation. As $\varepsilon \to 0$, $\mathbf{P}(t_i^+)$ will become singular, unlike the previous characterization of the algorithm. The result of the limiting process is

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i)]^{-1}[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \tag{5-126}$$

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i)]^{-1}\mathbf{H}(t_i)\mathbf{P}(t_i^-) \tag{5-127}$$

Let us investigate the invertibility of $[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i)]$ and the singularity of $\mathbf{P}(t_i^+)$ further. Let $\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i)$ be written out as

$$\left[\begin{array}{c} z_1(t_i) \\ \vdots \\ z_m(t_i) \end{array} \right] = \left[\begin{array}{c} \mathbf{h}_1^T(t_i) \\ \vdots \\ \mathbf{h}_m^T(t_i) \end{array} \right] \left[\begin{array}{c} x_1(t_i) \\ \vdots \\ x_n(t_i) \end{array} \right] = \left[\begin{array}{c} \mathbf{h}_1^T(t_i)\mathbf{x}(t_i) \\ \vdots \\ \mathbf{h}_m^T(t_i)\mathbf{x}(t_i) \end{array} \right]$$

Thus, if $H(t_i)$ is of full rank [i.e., if $\mathbf{h}_1(t_i), \ldots, \mathbf{h}_m(t_i)$ are a linearly independent set of m vectors], then there are m directions in state space along which we have perfect measurements (if it is not of full rank, then there are rank $[H(t_i)]$ such directions). With respect to some coordinate frame in state space, then, we can get perfect measurements of m (rank $[H(t_i)]$) out of the n states. Figure 5.13 depicts the case of a two-dimensional measurement and a three-dimensional state: by knowing z_1 and z_2, we know the state in two of three directions perfectly, i.e., in a two-dimensional subspace of R^3.
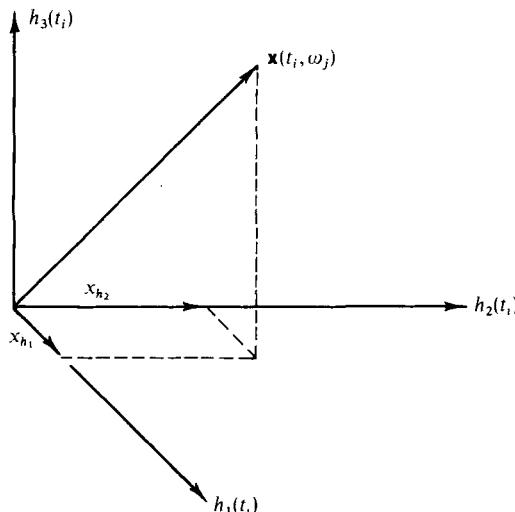
FIG. 5.13   Perfect two-dimensional measurement in three-dimensional state space. $x_{h_1} = (1/\|\mathbf{h}_1(t_i)\|)z_1$; $x_{h_2} = (1/\|\mathbf{h}_2(t_i)\|)z_2$.

If $P(t_i^-)$ is positive definite (of rank n) and $H(t_i)$ is of full rank m, then $[H(t_i)P(t_i^-)H^T(t_i)]$ is a positive definite m-by-m matrix, and so has an inverse. $P(t_i^+)$ will then be singular, of rank $(n - m)$; m eigenvalues are zero. This singularity is readily seen by premultiplying (5-127) by $H(t_i)$ to obtain

$$H(t_i)P(t_i^+) = H(t_i)P(t_i^-) - H(t_i)P(t_i^-)H^T(t_i)[H(t_i)P(t_i^-)H^T(t_i)]^{-1}H(t_i)P(t_i^-)$$
$$= H(t_i)P(t_i^-) - H(t_i)P(t_i^-) = 0 \qquad (5\text{-}128)$$

Heuristically, when you take a perfect m-vector measurement, the error probability density collapses in the directions along which you can determine the values of the state components exactly.

Thus, we want $P(t_i^-)$ to be positive definite (of rank n), whereas $P(t_{i-1}^+)$ is singular (positive semidefinite, of rank $n - m$), with

$$P(t_i^-) = \Phi(t_i, t_{i-1})P(t_{i-1}^+)\Phi^T(t_i, t_{i-1}) + \int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)G(\tau)Q(\tau)G^T(\tau)\Phi^T(t_i, \tau)\,d\tau$$

One *sufficient* condition for this to be true would be for the integral term itself to be of rank n: for the system representation to be stochastically controllable over each sample period.

EXAMPLE 5.11   Consider the gyro example again, but now let the measurement be corrupted by a bias only, with no additional white noise corruption, as depicted in Fig. 5.14. The bias is modeled as the output of an undriven integrator: the shaping filter is simply $\dot{x}_f(t) = 0$. In augmented state vector form, this becomes

$$\begin{bmatrix} \dot{x}(t) \\ \dot{x}_f(t) \end{bmatrix} = \begin{bmatrix} -\alpha & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ x_f(t) \end{bmatrix} + \begin{bmatrix} \alpha \\ 0 \end{bmatrix} w(t)$$

$$z(t_i) = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x(t_i) \\ x_f(t_i) \end{bmatrix}$$

The initial conditions are assumed to be

$$E\{x(t_0)\} = E\{x_f(t_0)\} = 0$$

$$E\{x^2(t_0)\} = E\{x_f{}^2(t_0)\} = 1 \qquad E\{x(t_0)x_f(t_0)\} = 0$$

so the appropriate estimator initial conditions are

$$\hat{x}(t_0) = \hat{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad P(t_0) = P_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

For the given augmented F_a, the state transition matrix is

$$\Phi_a(t_i, t_{i-1}) = \begin{bmatrix} e^{-\alpha(t_i - t_{i-1})} & 0 \\ 0 & 1 \end{bmatrix} \cong \begin{bmatrix} 0.78 & 0 \\ 0 & 1 \end{bmatrix}$$

The time propagation relations are, by (5-36) and (5-37),

$$\begin{bmatrix} \hat{x}(t_i^-) \\ \hat{x}_f(t_i^-) \end{bmatrix} = \begin{bmatrix} 0.78 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}(t_{i-1}^+) \\ \hat{x}_f(t_{i-1}^+) \end{bmatrix} = \begin{bmatrix} 0.78\hat{x}(t_{i-1}^+) \\ \hat{x}_f(t_{i-1}^+) \end{bmatrix}$$

$$P(t_i^-) = \Phi(t_i, t_{i-1})P(t_{i-1}^+)\Phi^T(t_i, t_{i-1}) + \int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)GQG^T\Phi^T(t_i, \tau)\,d\tau$$
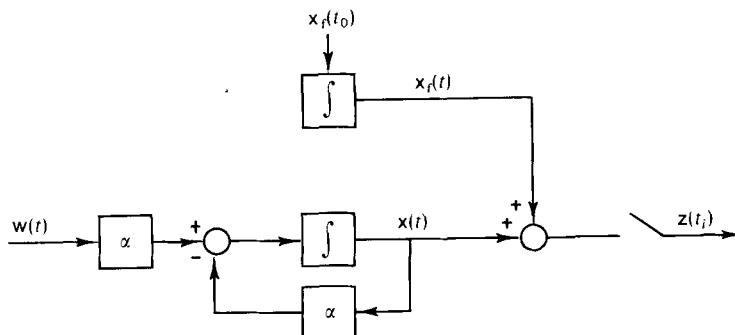
FIG. 5.14   Gyro corrupted by a bias.

The measurement update relations are given by (5-126) and (5-127) as

$$\begin{bmatrix} K_1(t_i) \\ K_2(t_i) \end{bmatrix} = \begin{bmatrix} [P_{11}(t_i^-) + P_{12}(t_i^-)]/[P_{11}(t_i^-) + 2P_{12}(t_i^-) + P_{22}(t_i^-)] \\ [P_{12}(t_i^-) + P_{22}(t_i^-)]/[P_{11}(t_i^-) + 2P_{12}(t_i^-) + P_{22}(t_i^-)] \end{bmatrix}$$

$$\begin{bmatrix} \hat{x}(t_i^+) \\ \hat{x}_f(t_i^+) \end{bmatrix} = \begin{bmatrix} \hat{x}(t_i^-) \\ \hat{x}_f(t_i^-) \end{bmatrix} + \begin{bmatrix} K_1(t_i) \\ K_2(t_i) \end{bmatrix} [z_i - \hat{x}(t_i^-) - \hat{x}_f(t_i^-)]$$

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{K}(t_i)\mathbf{H}\mathbf{P}(t_i^-)$$

By direct addition of the above expressions, it can be seen that

$$\hat{x}(t_i^+) + \hat{x}_f(t_i^+) = z_i$$

Such a phenomenon invariably occurs due to the perfect measurement of $[x(t_i) + x_f(t_i)]$. As shown in Fig. 5.15, when a measurement $z(t_i, \omega_j) = z_i$ becomes available, we know the value of $[x(t_i, \omega_j) + x_f(t_i, \omega_j)]$ exactly: the probability density describing the possible values of $x(t_i)$ and $x_f(t_i)$ has collapsed down to being nonzero only above the line $[x(t_i, \omega_j) + x_f(t_i, \omega_j)] = z_i$. (As time goes on between measurements, the probability density spreads out again.) Thus, $\mathbf{P}(t_i^+)$ is singular, of rank one. If we were to rotate coordinates (by similarity transformation) to the $\xi_1^* - \xi_2^*$ co-ordinates in the figure, there would be no uncertainty in the ξ_1^* direction just after a measurement, and so

$$\mathbf{P}^*(t_i^+) = \begin{bmatrix} P_{11}^*(t_i^+) & P_{12}^*(t_i^+) \\ P_{12}^*(t_i^+) & P_{22}^*(t_i^+) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & P_{22}^*(t_i^+) \end{bmatrix} \quad \blacksquare$$
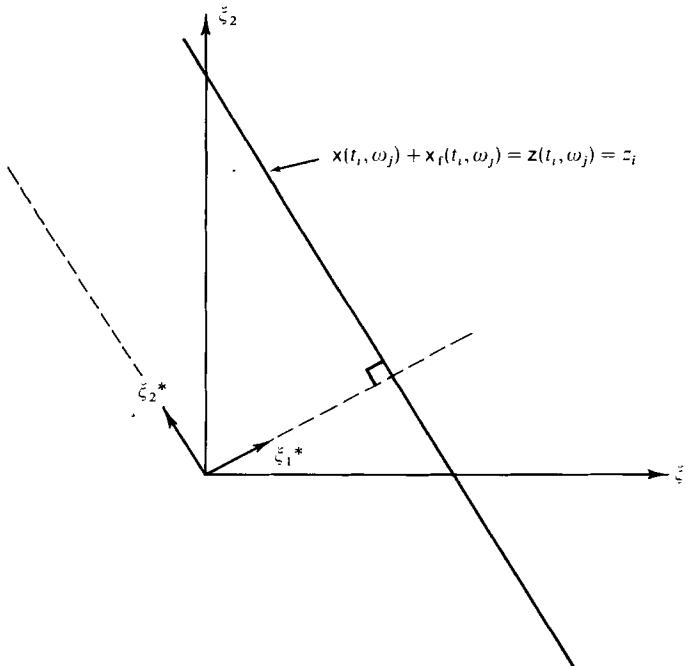
FIG. 5.15   Perfect measurement for Example 5.11.

As suggested in the example, it is convenient to define a coordinate transformation such that the perfect measurements are in fact direct measurements of individual state variables. Actually, a more general transformation than a simple rotation can be exploited to advantage. Assume $\mathbf{H}(t_i)$ to be full rank: this is not restrictive, since if it were not, we would have redundant perfect information. An n-by-n nonsingular transformation matrix $\mathbf{T}(t_i)$ can be defined for each $t_i \in T$ as

$$\mathbf{T}(t_i) = \left[\frac{\mathbf{H}(t_i)}{\mathbf{J}(t_i)}\right] \qquad (5\text{-}129)$$

where $\mathbf{H}(t_i)$ is the measurement matrix and $\mathbf{J}(t_i)$ is any convenient $(n - m)$-by-n matrix that yields a nonsingular (and thus invertible) $\mathbf{T}(t_i)$. Since this procedure is computationally attractive for the time-invariant $\mathbf{H}$ case, we confine our attention to

$$\mathbf{T}(t_i) = \mathbf{T} = \left[\frac{\mathbf{H}}{\mathbf{J}}\right] \qquad (5\text{-}129')$$

Through this transformation, the problem can be expressed in terms of $\mathbf{x}^*(t)$, where

$$\mathbf{x}^*(t) = \mathbf{T}\mathbf{x}(t), \qquad \mathbf{x}(t) = \mathbf{T}^{-1}\mathbf{x}^*(t) \qquad (5\text{-}130)$$

as the equivalent

$$\dot{\mathbf{x}}^*(t) = \mathbf{F}^*(t)\mathbf{x}^*(t) + \mathbf{B}^*(t)\mathbf{u}(t) + \mathbf{G}^*(t)\mathbf{w}(t), \qquad \mathbf{z}(t) = \mathbf{H}^*\mathbf{x}^*(t) \quad (5\text{-}131)$$

with $\mathbf{F}^*(t) = \mathbf{T}\mathbf{F}(t)\mathbf{T}^{-1}$, $\mathbf{B}^*(t) \doteq \mathbf{T}\mathbf{B}(t)$, and $\mathbf{G}^*(t) = \mathbf{T}\mathbf{G}(t)$. In view of (5-130),

$$\mathbf{x}^*(t_i) = \mathbf{T}\mathbf{x}(t_i) = \left[\frac{\mathbf{H}}{\mathbf{J}}\right]\mathbf{x}(t_i) = \left[\frac{\mathbf{z}(t_i)}{\mathbf{y}(t_i)}\right] \qquad (5\text{-}132)$$

In other words, the first m components of $\mathbf{x}^*(t_i)$ are just $\mathbf{z}(t_i)$, so $\mathbf{H}^* = [\mathbf{I} \mid \mathbf{0}]$, where $\mathbf{I}$ is m-by-m and the zero matrix is m-by-$(n - m)$. The remaining $(n - m)$ components of $\mathbf{x}^*(t_i)$ are identified as $\mathbf{y}(t_i)$: by choice of $\mathbf{J}$, this vector can often be a vector of variables to be estimated.

The optimal estimate of $\mathbf{x}^*(t_i)$ is obtained by operating on the measurement vector $\mathbf{z}_i$. The advantage in using the transformed variables is that the first m rows and columns of $\mathbf{P}(t_i^+)$ are identically zero and need not be computed. If desired, the optimal estimate of the original state vector $\mathbf{x}(t_i)$ can then be computed as $\hat{\mathbf{x}}(t_i^+) = \mathbf{T}^{-1}\hat{\mathbf{x}}^*(t_i^+)$.

EXAMPLE 5.12   In Example 5.11, the suggested coordinate rotation can be accomplished by

$$\mathbf{T} = \left[\frac{1 \quad 1}{-1 \quad 1}\right]$$

But, since we are interested specifically in obtaining an estimate of the first component of $\mathbf{x}_a(t_i)$, a more convenient choice of $\mathbf{J}$ would be $[1 \ 0]$, yielding

$$\mathbf{T} = \begin{bmatrix} 1 & 1 \\ \hline 1 & 0 \end{bmatrix}$$

Thus we get

$$\begin{bmatrix} \mathbf{x}_1{}^*(t_i) \\ \mathbf{x}_2{}^*(t_i) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_i) \\ \mathbf{x}_f(t_i) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(t_i) + \mathbf{x}_f(t_i) \\ \mathbf{x}(t_i) \end{bmatrix} = \begin{bmatrix} \mathbf{z}(t_i) \\ \mathbf{x}(t_i) \end{bmatrix}$$

Thus, the $\mathbf{y}(t_i)$ of (5-132) is in fact $\mathbf{x}(t_i)$, the variable to be estimated. The defining matrices are

$$\mathbf{F}^* = \mathbf{T}\mathbf{F}\mathbf{T}^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -\alpha & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -\alpha \\ 0 & -\alpha \end{bmatrix}$$

$$\mathbf{G}^* = \mathbf{T}\mathbf{G} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}$$

$$\mathbf{H}^* = [1 \ 0]$$

The filter update relations are

$$\mathbf{K}^*(t_i) = \mathbf{P}^*(t_i{}^-)\mathbf{H}^{*\mathrm{T}}[\mathbf{H}^*\mathbf{P}^*(t_i{}^-)\mathbf{H}^{*\mathrm{T}}]^{-1} = \begin{bmatrix} P_{11}^*(t_i{}^-)/P_{11}^*(t_i{}^-) \\ P_{12}^*(t_i{}^-)/P_{11}^*(t_i{}^-) \end{bmatrix} = \begin{bmatrix} 1 \\ P_{12}^*(t_i{}^-)/P_{11}^*(t_i{}^-) \end{bmatrix}$$

$$\mathbf{P}^*(t_i{}^+) = \begin{bmatrix} 0 & 0 \\ 0 & P_{22}^*(t_i{}^-) - [P_{12}^*(t_i{}^-)]^2/P_{11}^*(t_i{}^-)] \end{bmatrix}$$

In higher-dimensioned problems, there would be greater computational advantage to using transformed variables. ■

The construction of a state estimate given perfect measurements from a system that has no dynamic driving noise (but initial conditions are not known perfectly) can be developed by means of *Luenberger observers* [45, 46, 48]. Coordinate transformations similar to that of (5-129) are utilized in the design of observers, in which the state estimate is generated as $\hat{\mathbf{x}}(t_i{}^+) = \mathbf{T}^{-1}\hat{\mathbf{x}}^*(t_i)$, where $\hat{\mathbf{x}}^*(t_i)$ is given by (5-132), with $\mathbf{z}(t_i)$ provided by the measuring devices and $\mathbf{y}(t_i)$ the output of an $(n - m)$-dimensional linear system (the "minimal order observer"). The inherent freedom of choice of $\mathbf{J}$ is exploited to achieve desirable dynamic performance of the algorithm. *Observer-estimators* [78, 79] can be developed for the problem of state reconstruction in which noises drive *some* states and corrupt *some* measurements. This is of practical significance in cases involving large differences in precision of sensors so that, with respect to the finite wordlength of the digital computer being used, some measurements "look perfect" compared to others. Observer theory has also been extended to the stochastic case [42, 80, 81]. For details, see the cited references.

The problem of time-correlated measurements can be handled by an alternate approach [10] that avoids state augmentation (and thus increased state dimension). Instead, consecutive measurements are differenced to generate pseudo-measurements in which the corrupting noise is white. Consider the

discrete-time system and shaping filter representations

$$\mathbf{x}(t_{i+1}) = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \mathbf{G}_{\mathrm{d}}(t_i)\mathbf{w}_{\mathrm{d}}(t_i) \qquad (5\text{-}133\mathrm{a})$$

$$\mathbf{x}_{\mathrm{f}}(t_{i+1}) = \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{x}_{\mathrm{f}}(t_i) + \mathbf{G}_{\mathrm{df}}(t_i)\mathbf{w}_{\mathrm{df}}(t_i) \qquad (5\text{-}133\mathrm{b})$$

with $\mathbf{w}_{\mathrm{d}}(\cdot, \cdot)$ and $\mathbf{w}_{\mathrm{df}}(\cdot, \cdot)$ zero-mean independent white Gaussian noises of strengths $\mathbf{Q}_{\mathrm{d}}(t_i)$ and $\mathbf{Q}_{\mathrm{df}}(t_i)$, respectively, for all $t_i \in T$, and assume that $\mathbf{x}(t_0)$ and $\mathbf{x}_{\mathrm{f}}(t_0)$ are uncorrelated. Further assume the measurement to be of the form

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{x}_{\mathrm{f}}(t_i) \qquad (5\text{-}133\mathrm{c})$$

Note that this implies $\mathbf{x}_{\mathrm{f}}(t_i)$ is an m-vector process and that $\mathbf{H}_{\mathrm{f}}(t_i) \triangleq \mathbf{I}$ for all t_i: a definite restriction, except that the useful case of exponentially time-correlated noise on each scalar measurement fits this description. Now define the pseudo-measurement process $\mathbf{z}_{\mathrm{d}}(\cdot, \cdot)$, with d denoting difference, as

$$\mathbf{z}_{\mathrm{d}}(t_i) = \mathbf{z}(t_{i+1}) - \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{z}(t_i) \qquad (5\text{-}134)$$

Although $\mathbf{z}_{\mathrm{d}}(t_i, \omega_j)$ will yield information about $\mathbf{x}(t_i)$, it does not become available until one sample period later, since it requires knowledge of $\mathbf{z}(t_{i+1}, \omega_j)$. Writing $\mathbf{z}_{\mathrm{d}}(t_i)$ in terms of (5-133) yields

$$\begin{aligned}
\mathbf{z}_{\mathrm{d}}(t_i) &= \mathbf{z}(t_{i+1}) - \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{z}(t_i) \\
&= \left[\mathbf{H}(t_{i+1})\mathbf{x}(t_{i+1}) + \mathbf{x}_{\mathrm{f}}(t_{i+1})\right] \\
&\quad - \left[\boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{H}(t_i)\mathbf{x}(t_i) + \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{x}_{\mathrm{f}}(t_i)\right] \\
&= \mathbf{H}(t_{i+1})\boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}(t_i) + \mathbf{H}(t_{i+1})\mathbf{G}_{\mathrm{d}}(t_i)\mathbf{w}_{\mathrm{d}}(t_i) \\
&\quad + \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{x}_{\mathrm{f}}(t_i) + \mathbf{G}_{\mathrm{df}}(t_i)\mathbf{w}_{\mathrm{df}}(t_i) \\
&\quad - \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{H}(t_i)\mathbf{x}(t_i) - \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{x}_{\mathrm{f}}(t_i) \\
&= \left[\mathbf{H}(t_{i+1})\boldsymbol{\Phi}(t_{i+1}, t_i) - \boldsymbol{\Phi}_{\mathrm{f}}(t_{i+1}, t_i)\mathbf{H}(t_i)\right]\mathbf{x}(t_i) \\
&\quad + \left[\mathbf{H}(t_{i+1})\mathbf{G}_{\mathrm{d}}(t_i)\mathbf{w}_{\mathrm{d}}(t_i) + \mathbf{G}_{\mathrm{df}}(t_i)\mathbf{w}_{\mathrm{df}}(t_i)\right] \qquad (5\text{-}135)
\end{aligned}$$

This is in the form of a linear combination of the original system states, corrupted by a zero-mean white Gaussian noise of strength $\mathbf{R}_{\mathrm{d}}(t_i)$,

$$\begin{aligned}
\mathbf{R}_{\mathrm{d}}(t_i) &= \mathbf{H}(t_{i+1})\mathbf{G}_{\mathrm{d}}(t_i)\mathbf{Q}_{\mathrm{d}}(t_i)\mathbf{G}_{\mathrm{d}}^{\mathrm{T}}(t_i)\mathbf{H}^{\mathrm{T}}(t_{i+1}) \\
&\quad + \mathbf{G}_{\mathrm{df}}(t_i)\mathbf{Q}_{\mathrm{df}}(t_i)\mathbf{G}_{\mathrm{df}}^{\mathrm{T}}(t_i) \qquad (5\text{-}136)
\end{aligned}$$

Since $\mathbf{z}(t_{i+1}, \omega_j)$ is required before $\mathbf{z}_{\mathrm{d}}(t_i)$ can be processed, this actually yields an optimal smoothing problem formulation rather than an optimal filtering problem, and the algorithm can be developed using the results of Chapter 8 (Volume 2). In application, the measurement $\mathbf{z}(t_{i+1}, \omega_j)$ is taken at time t_{i+1}, $\mathbf{z}_{\mathrm{d}}(t_i, \omega_j)$ is thereby computed, $\hat{\mathbf{x}}(t_i)$ is then calculated, and the estimate propagated to the current real time t_{i+1} as $\hat{\mathbf{x}}(t_{i+1}^-)$.

5.11 CONTINUOUS-TIME FILTER

As discussed previously, practical application of optimal estimation almost invariably involves implementation on a digital computer, which inherently dictates sampled-data format for measurements. Consequently, attention has been concentrated on this formulation. This section provides a formal derivation of the continuous-data Kalman filter [15, 37]; it can be made rigorous, and the concepts involved are of considerable theoretical significance, but the additional difficulty and effort is not warranted in view of our objective to attain efficient, practical algorithms. One might consider generating the continuous filter and then discretizing it for eventual implementation, but there is a serious drawback to this procedure, discussed subsequently.

Consider the same continuous-time dynamics model as used before

$$\mathbf{dx}(t) = \mathbf{F}(t)\mathbf{x}(t)\,dt + \mathbf{B}(t)\mathbf{u}(t)\,dt + \mathbf{G}(t)\mathbf{d\beta}(t) \qquad (5\text{-}137)$$

or, in white noise notation

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t) \qquad (5\text{-}137')$$

where $\beta(\cdot,\cdot)$ is Brownian motion of diffusion $\mathbf{Q}(t)$ for all $t \in T$, or $\mathbf{w}(\cdot,\cdot)$ is zero-mean white Gaussian noise of strength $\mathbf{Q}(t)$ for all $t \in T$. Let $\mathbf{x}(t_0)$ be modeled as a Gaussian random variable with mean $\hat{\mathbf{x}}_0$ and covariance $\mathbf{P}_0$.

We want to consider continuously available measurements, modeled by the process $\mathbf{z}_c(\cdot,\cdot)$ defined by

$$\mathbf{z}_c(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}_c(t) \qquad (5\text{-}138)$$

where $\mathbf{v}_c(\cdot,\cdot)$ is a zero-mean white Gaussian noise with

$$E\{\mathbf{v}_c(t)\mathbf{v}_c{}^{\mathrm{T}}(t + \tau)\} = \mathbf{R}_c(t)\,\delta(\tau) \qquad (5\text{-}139)$$

The subscript c is meant to distinguish these continuous time processes from the analogous discrete-time processes considered previously.

To derive the desired result, we will consider a discrete-time measurement process and examine the result of letting the time between sample times decrease in the limit to zero. Thus, the measurements are described by

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \qquad (5\text{-}140)$$

where $\mathbf{v}(\cdot,\cdot)$ is a zero-mean white Gaussian sequence with

$$E\{\mathbf{v}(t_i)\mathbf{v}^{\mathrm{T}}(t_i)\} = \mathbf{R}(t_i) = \mathbf{R}_c(t_i)/\Delta t_i \qquad (5\text{-}141a)$$

$$E\{\mathbf{v}(t_i)\mathbf{v}^{\mathrm{T}}(t_j)\} = \mathbf{0}, \qquad i \neq j \qquad (5\text{-}141b)$$

where $\mathbf{R}(t_i)$ and $\mathbf{R}_c(t_i)$ are positive definite and symmetric for all $t_i \in T$ and Δt_i is the time interval $[t_{i+1} - t_i]$. Without any real loss of generality, we let all Δt_i's be the same, Δt, in the derivation. The covariance $\mathbf{R}_c(t_i)$ in (5-141a) will eventually become the strength of the continuous-time white Gaussian noise

process $\mathbf{v}_c(\cdot, \cdot)$ that corrupts the continuous-time measurements. This description results in an autocorrelation function as depicted in Fig. 5.16, defined for discrete values of τ. Note that as $\Delta t \to 0$, a Dirac delta function (an infinite impulse at $\tau = 0$) is achieved. We further assume that $\mathbf{x}(t_0)$, $\boldsymbol{\beta}(\cdot, \cdot)$ or $\mathbf{w}(\cdot, \cdot)$, and $\mathbf{v}(\cdot, \cdot)$ are independent, and that $\mathbf{F}(\cdot)$, $\mathbf{B}(\cdot)$, $\mathbf{G}(\cdot)$, $\mathbf{H}(\cdot)$, $\mathbf{Q}(\cdot)$, and $\mathbf{R}_c(\cdot)$ are at least piecewise continuous.
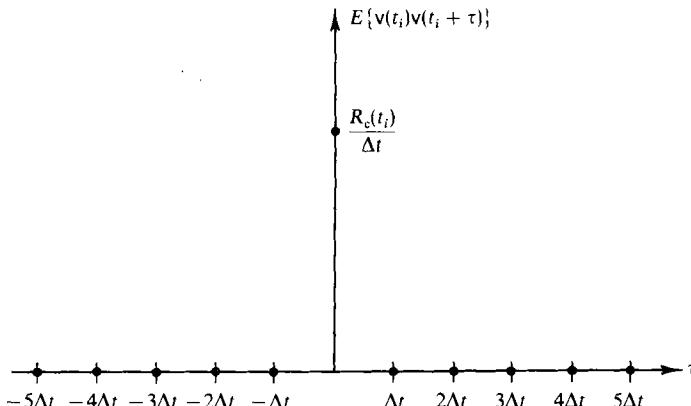
FIG. 5.16   Autocorrelation function for discrete-time $\mathbf{v}(\cdot, \cdot)$.

The discrete-time Kalman filter for this problem formulation is given by Eqs. (5-36)–(5-42) with $\mathbf{R}(t_i)$ repeated by $\mathbf{R}_c(t_i)/\Delta t$. In the time propagation relations, the state transition matrix can be expanded as

$$\boldsymbol{\Phi}(t_i, t_{i-1}) = \mathbf{I} + \mathbf{F}(t_{i-1})\Delta t + \mathcal{O}(\Delta t^2) \tag{5-142}$$

where $\mathcal{O}(\Delta t^2)$ is composed of terms involving powers of Δt greater than or equal to two, such that

$$\lim_{\Delta t \to 0} \frac{\mathcal{O}(\Delta t^2)}{\Delta t} = \mathbf{0}$$

First consider the state estimate equations: substituting (5-142) into the $\hat{\mathbf{x}}(t_i^-)$ equation, (5-36), yields

$$\hat{\mathbf{x}}(t_i^-) = [\mathbf{I} + \mathbf{F}(t_{i-1})\Delta t]\hat{\mathbf{x}}(t_{i-1}^+) + \int_{t_{i-1}}^{t_i} [\mathbf{I} + \mathbf{F}(\tau)\{t_i - \tau\}]\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau + \mathcal{O}(\Delta t^2)$$

$$= \hat{\mathbf{x}}(t_{i-1}^+) + \mathbf{F}(t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+)\Delta t + \mathbf{B}(\sigma)\mathbf{u}(\sigma)\Delta t + \mathcal{O}(\Delta t^2) \tag{5-143}$$

where σ is somewhere in the interval $[t_{i-1}, t_i)$, by the mean value theorem. Substituting (5-143) into the $\hat{\mathbf{x}}(t_i^+)$ equation, (5-39), yields

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_{i-1}^+) + \mathbf{F}(t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+)\Delta t + \mathbf{B}(\sigma)\mathbf{u}(\sigma)\Delta t + \mathcal{O}(\Delta t^2)$$
$$+ \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i)\Delta t + \mathbf{R}_c(t_i)]^{-1}\Delta t[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$$

Now when $\hat{\mathbf{x}}(t_{i-1}^+)$ is brought to the left hand side of this equation, the entire result divided by Δt, and the limit taken as $\Delta t \to 0$, we get

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{F}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{P}(t)\mathbf{H}^{\mathrm{T}}(t)\mathbf{R}_c^{-1}(t)[\mathbf{z}(t) - \mathbf{H}(t)\hat{\mathbf{x}}(t)] \quad (5\text{-}144)$$

where, in the limit, $\hat{\mathbf{x}}(t_{i-1}^+) \to \hat{\mathbf{x}}(t_i^-) \to \hat{\mathbf{x}}(t_i^+) \triangleq \hat{\mathbf{x}}(t)$ and $\mathbf{P}(t_{i-1}^+) \to \mathbf{P}(t_i^-) \to \mathbf{P}(t_i^+) \triangleq \mathbf{P}(t)$. Note that $\mathbf{z}(\cdot)$ is a sample from the continuous-time measurement process $\mathbf{z}_c(\cdot,\cdot)$. Performing similar operations on the covariance equations yields

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) - \mathbf{P}(t)\mathbf{H}^{\mathrm{T}}(t)\mathbf{R}_c^{-1}(t)\mathbf{H}(t)\mathbf{P}(t)$$

$$(5\text{-}145)$$

This has not been very rigorous, but existence of the desired limits can be proven by means of probability one arguments and the concept of martingales, to prove, among other things, that

$$\lim_{k \to \infty} E\{\mathbf{x}(t) \mid \mathbf{z}(t_1) = \mathbf{z}_1, \ldots, \mathbf{z}(t_k) = \mathbf{z}_k; t_1, \ldots, t_k \le t\}$$

$$= E\{\mathbf{x}(t) \mid \mathbf{z}_c(\tau) = \mathbf{z}(\tau); t_0 \le \tau \le t\}$$

We will gloss over these aspects here. Thus, the continuous-time Kalman filter is specified by the differential equations (5-144) and (5-145), which are integrated forward from the initial conditions

$$\hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0 \qquad (5\text{-}146a)$$

$$\mathbf{P}(t_0) = \mathbf{P}_0 \qquad (5\text{-}146b)$$

Figure 5.17 portrays the basic system model and the continuous-time Kalman filter based upon this model. From this figure it is evident that within the structure of the filter is a mathematical model of the real system that provides the measurement data input to the filter, just as in the discrete-time measurement case. The filter incorporates such a model, driven by an optimal gain times the difference between the actual measurements received, $\mathbf{z}(t)$, and the optimal estimates of what these should be based on the mathematical model output, $\mathbf{H}(t)\hat{\mathbf{x}}(t)$, the residual $[\mathbf{z}(t) - \mathbf{H}(t)\hat{\mathbf{x}}(t)]$.

In the continuous-time case, the Kalman filter gain is seen to be

$$\mathbf{K}(t) = \mathbf{P}(t)\mathbf{H}^{\mathrm{T}}(t)\mathbf{R}_c^{-1}(t) \qquad (5\text{-}147)$$

If $\mathbf{P}(t)$ is "large" (having large eigenvalues), then the residual is heavily weighted: if we are very uncertain of the current estimate $\hat{\mathbf{x}}(t)$, then the new information from the measurements is emphasized. Similarly, if $\mathbf{R}_c(t)$ is "small," i.e., if the measurements are very accurate, then the measurement information is weighted heavily. In fact, $\mathbf{P}(t)\mathbf{H}^{\mathrm{T}}(t)\mathbf{R}_c^{-1}(t)$ can be interpreted heuristically as a signal to noise ratio.
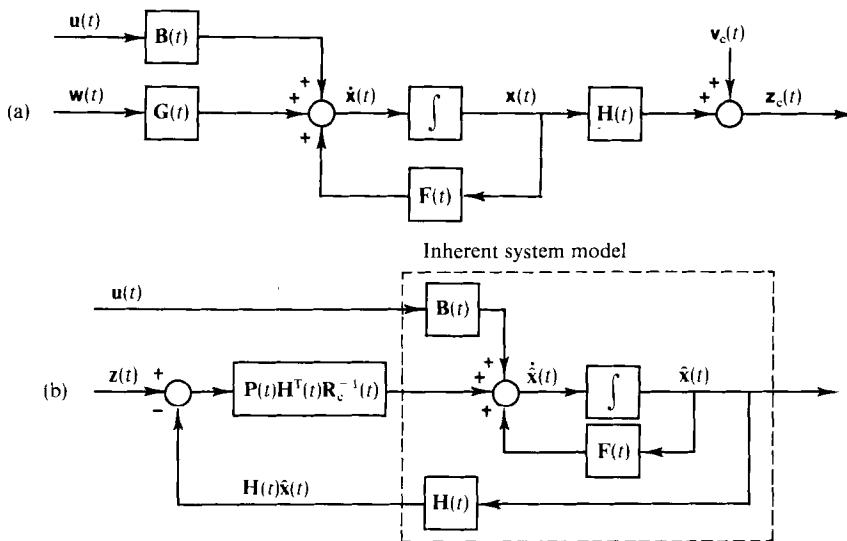
FIG. 5.17   (a) Continuous-time system model and (b) Kalman filter.

An algebraically equivalent form for the error covariance equation, (5-145), in terms of the Kalman gain $\mathbf{K}(t)$ of (5-147), is

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) - \mathbf{K}(t)\mathbf{R}_{\mathrm{c}}(t)\mathbf{K}^{\mathrm{T}}(t) \quad (5\text{-}148)$$

The first two terms of either of these expressions indicate the homogeneous system effects (usually stabilizing), the third term is the covariance increasing effect due to the dynamic noise $\mathbf{w}(\cdot,\cdot)$, and the fourth term is the error covariance decreasing effect of incorporating the measurement information. Thus it is reasonable that removing the last term, i.e., removing the continuously available measurements, yields the relation for propagating the error covariance between sample times in the sampled-data Kalman filter.

Equation (5-145) or (5-148) is a continuous-time matrix Riccati differential equation. As a regular differential equation, it has the usual properties of existence, uniqueness, and continuity of solutions. However, Riccati equations are very difficult to integrate numerically, being very sensitive to integration step size and often exhibiting unstable computed solutions despite theoretical solution stability.

Most practical estimation problems are characterized by system dynamics most naturally modeled by differential, rather than difference, equations. However, the designer knows from the outset that a digital computer will be employed in the eventual estimator implementation, thus dictating sampled-data rather than continuous-time measurements. Consequently, there are two means of systematic estimator design. First, we could take the continuous-time system model, design the continuous-time filter, and then discretize the result. Second, we could determine an equivalent discrete-time model and generate

the discrete-time filter from it. However, the first approach is fraught with the difficulties of integrating Riccati differential equations, while the second involves significantly better behaved recursions. Moreover, the discretization of a continuous filter is an approximation to an optimal discrete-time filter, whereas a discrete-time filter based on an *equivalent* discrete-time model involves no approximations. Thus, the preferable design approach is to discretize the model first and then generate the filter.

EXAMPLE 5.13 Once again we examine the gyro on test, but now assuming that measurement data are available continuously, modeled as

$$z_c(t) = x(t) + v_c(t)$$

where $v_c(\cdot, \cdot)$ is a zero-mean white Gaussian noise with $E\{v_c(t)v_c(t + \tau)\} = R_c(t)\delta(\tau)$. In view of (5-141a), if we want the continuous-time estimator performance to approximate that of the discrete-time filter, the appropriate noise strength would be

$$R_c(t) = R_c = [R(t_i)\Delta t] = (0.5 \quad \text{deg}^2/\text{hr}^2)(0.25 \quad \text{hr}) = 0.125 \quad \text{deg}^2/\text{hr}$$

The estimator is given by

$$\dot{\hat{x}}(t) = -\alpha\hat{x}(t) + [P(t)/R_c][z(t) - \hat{x}(t)] = -\hat{x}(t) + [8P(t)][z(t) - \hat{x}(t)]$$

integrated forward from the initial condition $\hat{x}(t_0) = 0$, where $P(t)$ satisfies the Riccati equation

$$\dot{P}(t) = -2\alpha P(t) + Q - [P^2(t)/R_c] = -2P(t) + 2 - 8P^2(t)$$

with an initial condition of $P(t_0) = 1$. For this problem, a steady state value of $P(t)$ is achieved, and it can be found by setting $\dot{P}(t) = 0$:

$$2 - 2P - 8P^2 = 0$$

for which the positive solution is

$$P = 0.392$$

The total solution for $P(t)$ is, letting $a = \sqrt{\alpha^2 + (Q/R_c)} = \sqrt{17}$

$$P(t) = \frac{a + [Q - \alpha P_0]\tanh(a[t - t_0])}{a + [\alpha - R_c^{-1}P_0]\tanh(a[t - t_0])} = \frac{\sqrt{17} + \tanh(\sqrt{17}[t - t_0])}{\sqrt{17} + 9\tanh(\sqrt{17}[t - t_0])}$$

Figure 5.18 plots this result, superimposed upon the result of Example 5.3. The continuous-time solution passes through the region bounded by the oscillations of the discrete-time filter. As $\Delta t \to 0$, these variations converge to the continuous-time result. For instance, consider halving the sample period. As a result, the increase in $P(t)$ is not as great before the next measurement becomes available. However, to maintain constant R_c, the discrete-time $R(t_i)$ would have to be doubled according to (5-141a) [we want to let $\Delta t \to 0$ while letting $v(t_i) \to v_c(t_i)$ of strength R_c]. Each decrease in $P(t)$ due to measurement updating is thus less, since the measurements are now corrupted by stronger noise.

Note that if the product $[R(t_i)\Delta t]$ were not held constant, the discrete-time solutions would not converge to the continuous-time solution. If $R(t_i)$ were kept constant instead, more of equally accurate data would be available by halving the sample period. In the limit, perfect continuous measurements would be achieved, and $P(t)$ would instantaneously go to zero at t_0 and stay there for all time. ■

The numerical characteristics of the Riccati differential equation solution, especially sensitivity to integration step size, motivate either avoiding methods
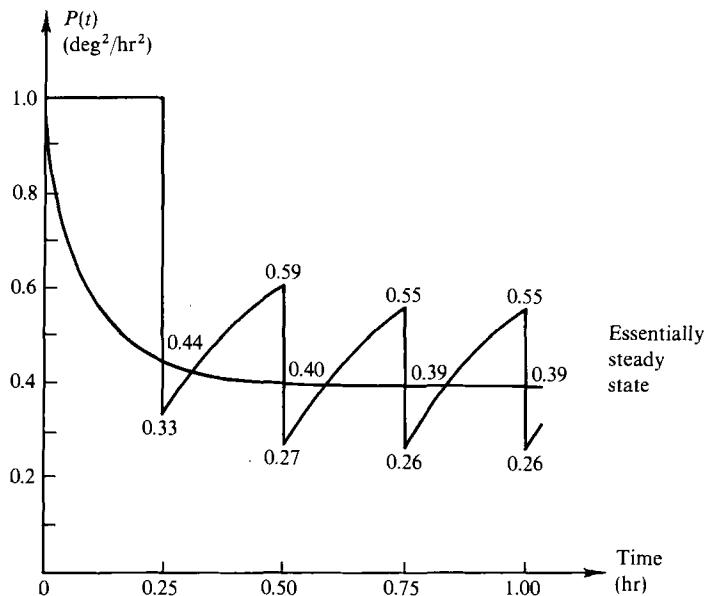
FIG. 5.18   Error variance of continuous-time filter for gyro example.

dependent on its solution or seeking solution techniques other than straight-forward integration. To generate a solution to an n-by-n Riccati matrix differential equation, it is possible to exploit an associated $2n$-by-n linear matrix differential equation. Specifically, a solution to (5-145) can be expressed as

$$\dot{P}(t) = U(t)V^{-1}(t) \tag{5-149}$$

where $U(t)$ and $V(t)$ are n-by-n matrices satisfying the homogeneous linear differential equation and initial condition

$$\begin{bmatrix} \dot{U}(t) \\ \hline \dot{V}(t) \end{bmatrix} = \begin{bmatrix} F(t) & \Big| & G(t)Q(t)G^T(t) \\ \hline H^T(t)R_c^{-1}(t)H(t) & \Big| & -F^T(t) \end{bmatrix} \begin{bmatrix} U(t) \\ \hline V(t) \end{bmatrix} \tag{5-150a}$$

$$\begin{bmatrix} U(t_0) \\ \hline V(t_0) \end{bmatrix} = \begin{bmatrix} P_0 \\ \hline I \end{bmatrix} \tag{5-150b}$$

where $V(t)$ is always invertible.

Proof of Equivalency   First, the assumed form (5-149) satisfies the initial condition

$$P(t_0) = U(t_0)V^{-1}(t_0) = P_0 I^{-1} = P_0$$

Differentiating the assumed form yields (dropping the time notation for convenience)

$$\dot{P} = \dot{U}V^{-1} + U\, d\{V^{-1}\}/dt$$

But, since $VV^{-1} = I, d\{VV^{-1}\}/dt = 0 = \dot{V}V^{-1} + V\, d\{V^{-1}\}/dt$, so

$$dV^{-1}/dt = -V^{-1}\dot{V}V^{-1}$$

if $\mathbf{V}(t)$ is in fact invertible, giving

$$\dot{\mathbf{P}} = \dot{\mathbf{U}}\mathbf{V}^{-1} - \mathbf{U}\mathbf{V}^{-1}\dot{\mathbf{V}}\mathbf{V}^{-1}$$

Substituting the partitions of (5-150a) into this yields

$$\dot{\mathbf{P}} = (\mathbf{F}\mathbf{U} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathsf{T}}\mathbf{V})\mathbf{V}^{-1} - \mathbf{U}\mathbf{V}^{-1}(\mathbf{H}^{\mathsf{T}}\mathbf{R}_{\mathrm{c}}^{-1}\mathbf{H}\mathbf{U} - \mathbf{F}^{\mathsf{T}}\mathbf{V})\mathbf{V}^{-1}$$

$$= \mathbf{F}\mathbf{U}\mathbf{V}^{-1} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathsf{T}}\mathbf{V}\mathbf{V}^{-1} - \mathbf{U}\mathbf{V}^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{R}_{\mathrm{c}}^{-1}\mathbf{H}\mathbf{U}\mathbf{V}^{-1} + \mathbf{U}\mathbf{V}^{-1}\mathbf{F}^{\mathsf{T}}\mathbf{V}\mathbf{V}^{-1}$$

$$= \mathbf{F}\mathbf{P} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathsf{T}} - \mathbf{P}\mathbf{H}^{\mathsf{T}}\mathbf{R}_{\mathrm{c}}^{-1}\mathbf{H}\mathbf{P} + \mathbf{P}\mathbf{F}^{\mathsf{T}}$$

which is in fact the original Riccati equation to be solved, (5-145). The matrix $\mathbf{V}(t)$ can be shown to be of full rank, and thus invertible, for all time. ■

Although (5-150) is a homogeneous linear differential equation, it embodies unstable modes; straightforward integration is not generally practicable, but eigenvalue techniques provide a useful means of attaining the steady state $\mathbf{P}$ satisfying (5-145) for the case of time-invariant systems with stationary noises [12, 58, 59]. Other means of solving (5-145) include iterative procedures [38–41], perturbation methods [61], the partitioned algorithm approach [43, 44], matrix factorization methods [63], the matrix sign function method [6, 19], and other means of enhancement in numerical integration [85].

The stability characteristics of the continuous-time filter are analogous to those of the discrete-time filter, described in Section 5.8. Here we consider the homogeneous portion of the filter state equations,

$$\dot{\hat{\mathbf{x}}}(t) = \left[\mathbf{F}(t) - \mathbf{P}(t)\mathbf{H}^{\mathsf{T}}(t)\mathbf{R}_{\mathrm{c}}^{-1}(t)\mathbf{H}(t)\right]\hat{\mathbf{x}}(t) + \mathbf{P}(t)\mathbf{H}^{\mathsf{T}}(t)\mathbf{R}_{\mathrm{c}}^{-1}(t)\mathbf{z}(t) \quad (5\text{-}151)$$

Then, if the system model upon which the continuous-time Kalman filter is based is stochastically observable (5-109) and stochastically controllable (5-105), then the filter is uniformly asymptotically globally stable. As in the sampled-data case, it is possible to generate a stable filter from an unstable system model.

Unlike the discrete-time case, $\mathbf{R}_{\mathrm{c}}^{-1}(t)$ appears explicitly in the continuous-time Kalman filter gain, (5-147), so that a singular $\mathbf{R}_{\mathrm{c}}(t)$ will require a substantial modification in the optimal estimator algorithm. Care must be taken when performing a limiting procedure to derive the optimal continuous-time estimator for problems characterized by time-correlated or no measurement noise, as considered in Section 5.10. (As in that section, filtering in the presence of only time-correlated measurement noise is a special case of filtering with perfect measurements, once state vector augmentation is exploited.) In the limit, the problem formulation and the optimal estimator, or *Deyst filter* [11, 13, 20, 21, 74], are as follows. Let the system dynamics be modeled as in (5-137) and let the continuous-time measurements be free of white noise corruption:

$$\mathbf{z}_{\mathrm{c}}(t) = \mathbf{H}(t)\mathbf{x}(t) \qquad (5\text{-}152)$$

Assume that $\mathbf{F}(\cdot)$, $\mathbf{B}(\cdot)$, $\mathbf{G}(\cdot)$, and $\mathbf{Q}(\cdot)$ are continuous with continuous first derivatives and $\mathbf{H}(\cdot)$ is continuous with continuous first and second derivatives.

Then the optimal estimator is specified in terms of an n-dimensional state $y(t)$ as

$$\hat{\mathbf{x}}(t) = \mathbf{y}(t) + \mathbf{W}(t)\mathbf{z}(t) \tag{5-153a}$$

$$\dot{\mathbf{y}}(t) = \{[\mathbf{I} - \mathbf{W}(t)\mathbf{H}(t)]\mathbf{F}(t) - \dot{\mathbf{W}}(t)\mathbf{H}(t) - \mathbf{W}(t)\dot{\mathbf{H}}(t)\}\hat{\mathbf{x}}(t)$$
$$\quad + [\mathbf{I} - \mathbf{W}(t)\mathbf{H}(t)]\mathbf{B}(t)\mathbf{u}(t) \tag{5-153b}$$

$$\mathbf{A}(t) = \mathbf{P}(t)\dot{\mathbf{H}}^{\mathrm{T}}(t) + [\mathbf{P}(t)\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t)]\mathbf{H}^{\mathrm{T}}(t) \tag{5-153c}$$

$$\mathbf{W}(t) = \mathbf{A}(t)[\mathbf{H}(t)\mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t)\mathbf{H}^{\mathrm{T}}(t)]^{-1} \tag{5-153d}$$

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) - \mathbf{W}(t)\mathbf{A}^{\mathrm{T}}(t) \tag{5-153e}$$

where (5-153a) and (5-153b) determine the basic filter structure, $\mathbf{W}(t)$ in (5-153d) is the filter weighting (gain) matrix, and (5-153e) is a matrix Riccati differential equation for the estimate error covariance. Figure 5.19 portrays the estimator structure. Note that, unlike the Kalman filter, the measurement $\mathbf{z}(t)$ can appear in the filter output *directly* with no integration. Because of this direct feed-through, in stationary-noise, time-invariant-system cases this filter can be described by a matrix of transfer functions, each of which are able to have a numerator of degree equal to that of the denominator. On the other hand, transfer functions corresponding to the steady state Kalman filter must have numerators of degree less than the corresponding denominators.
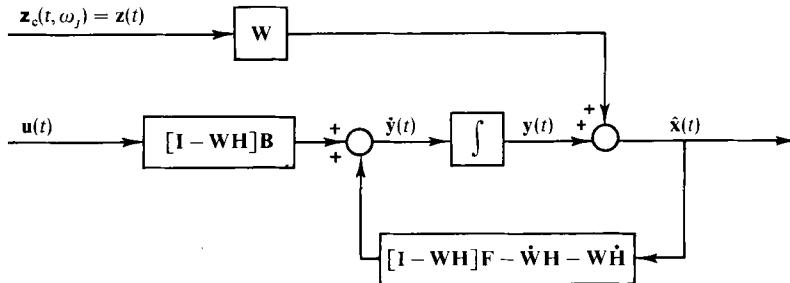
FIG. 5.19   Deyst filter structure.

Direct feedthrough of $\mathbf{z}(t)$ also affects the initialization of the filter. Before any measurement data are taken, the a priori statistics of $\mathbf{x}(t_0)$ are $\hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0$ and $\mathbf{P}(t_0) = \mathbf{P}_0$. However, at time t_0 there is a discontinuity [the form of which for $\hat{\mathbf{x}}(t_0^+)$ and $\mathbf{P}(t_0^+)$ is evident from a discrete-time Kalman filter update]:

$$\hat{\mathbf{x}}(t_0^+) = \hat{\mathbf{x}}_0 + \mathbf{P}_0\mathbf{H}^{\mathrm{T}}(t_0)[\mathbf{H}(t_0)\mathbf{P}_0\mathbf{H}^{\mathrm{T}}(t_0)]^{-1}[\mathbf{z}(t_0) - \mathbf{H}(t_0)\hat{\mathbf{x}}_0] \tag{5-154a}$$

$$\mathbf{P}(t_0^+) = \mathbf{P}_0 - \mathbf{P}_0\mathbf{H}^{\mathrm{T}}(t_0)[\mathbf{H}(t_0)\mathbf{P}_0\mathbf{H}^{\mathrm{T}}(t_0)]^{-1}\mathbf{H}(t_0)\mathbf{P}_0 \tag{5-154b}$$

$$\mathbf{A}(t_0^+) = \mathbf{P}(t_0^+)\dot{\mathbf{H}}^{\mathrm{T}}(t_0) + [\mathbf{P}(t_0^+)\mathbf{F}^{\mathrm{T}}(t_0) + \mathbf{G}(t_0)\mathbf{Q}(t_0)\mathbf{G}^{\mathrm{T}}(t_0)]\mathbf{H}^{\mathrm{T}}(t_0) \tag{5-154c}$$

$$\mathbf{W}(t_0^+) = \mathbf{A}(t_0^+)[\mathbf{H}(t_0)\mathbf{G}(t_0)\mathbf{Q}(t_0)\mathbf{G}^{\mathrm{T}}(t_0)\mathbf{H}^{\mathrm{T}}(t_0)]^{-1} \tag{5-154d}$$

$$\mathbf{y}(t_0^+) = \hat{\mathbf{x}}(t_0^+) - \mathbf{W}(t_0^+)\mathbf{z}(t_0) \tag{5-154e}$$

Thus, the filter cannot be initialized until the measurement value at time t_0 becomes available.

The m-by-m matrix $[H(t)G(t)Q(t)G^T(t)H^T(t)]$ must be nonsingular to be assured of the existence of its inverse, required by the algorithm in (5-153d). Unless $H(t)$ is of full rank for all $t \in T$, it will be singular; but linearly dependent rows of H can always be ignored since this is just redundant perfect information. Heuristically, this matrix represents the strength of first integrals of white noise in the measurements; its singularity implies that there exist one or more measurements, or linear combinations thereof, that contain no first integrals of white noise. If there is no white noise entering the system model just one integration before the measurement, one or more differentiators can be inserted into the filter input channel to generate derivatives of the measurements as "new" measurements (since we are more than one integration away from a white noise source, the differentiation process is not troublesome).

EXAMPLE 5.14   Consider the gyro with a bias as in Examples 5.11 and 5.12, with model depicted as in Fig. 5.14, but now assuming that measurements are available continuously. In terms of the transformed coordinates of Example 5.12, we have $[x_1(t) = z(t), x_2(t) = $ gyro drift rate to be estimated]

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & -\alpha \\ 0 & -\alpha \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} \alpha \\ \alpha \end{bmatrix} w(t)$$

$$z(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

To apply the Deyst filter equations, $[HGQG^TH^T]$ must be evaluated:

$$[HGQG^TH^T] = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \alpha \end{bmatrix} Q \begin{bmatrix} \alpha & \alpha \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \alpha^2 Q$$

This is invertible, as expected, since the measurement is separated from the white noise $w(\cdot,\cdot)$ by only one integration.

Consider the covariance. The initial condition is given in terms of the components of P_0 by (5-154b)

$$P(t_0{}^+) = \begin{bmatrix} 0 & 0 \\ 0 & P_{0_{22}} - [P_{0_{12}}^2/P_{0_{11}}] \end{bmatrix}$$

where the first row and column are zero because we know $z(t_0)$ exactly. The Riccati equation, (5-153e), becomes

$$\begin{bmatrix} \dot{P}_{11}(t) & \dot{P}_{12}(t) \\ \dot{P}_{12}(t) & \dot{P}_{22}(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -P_{22}(t)^2/Q \end{bmatrix}$$

again because $z(t)$ is known exactly. Thus, the only Riccati equation to solve is the scalar equation for $\dot{P}_{22}(t)$, for which the solution is

$$P_{22}(t) = [QP_{22}(t_0{}^+)]/[Q + P_{22}(t_0{}^+)\{t - t_0\}]$$

This reduction of the dimension of the Riccati matrix equation by m, the number of measurements, always occurs in the transformed coordinates. Note that as $t \to \infty$, $P_{22}(t) \to 0$.

Substituting into the rest of the filter equations yields

$$\hat{x}_1(t) = y_1(t) + z(t)$$
$$\hat{x}_2(t) = y_2(t) + [1 - P_{22}(t)/(\alpha Q)]z(t)$$

where

$$\dot{y}_1(t) = 0$$
$$\dot{y}_2(t) = -[P_{22}(t)/Q]\hat{x}_2(t) + [\dot{P}_{22}(t)/(\alpha Q)]z(t)$$

But it can be shown that $y_1(t_0{}^+) = 0$, so $y_1(t) = 0$ for all time t, and then

$$\hat{x}_1(t) = z(t)$$

as anticipated. A block diagram of the filter is given in Fig. 5.20. Note the direct feedthrough of $z(t)$ through C_1. In this filter, $\hat{x}_2(t)$ is in fact the estimate of the gyro drift rate.   ■
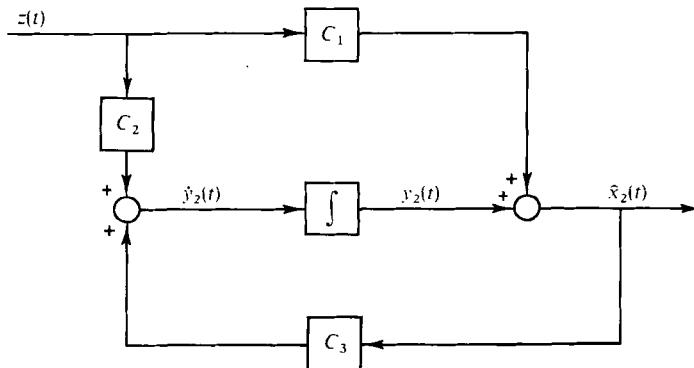
FIG. 5.20   Deyst filter for Example 5.14. $C_1 = 1 - [P_{22}(t)/\alpha Q]$, $C_2 = \dot{P}_{22}(t)/\alpha Q = -P_{22}(t)^2/\alpha Q^2$, $C_3 = -P_{22}(t)/Q$.

EXAMPLE 5.15   To see the effect of singular $[HGQG^TH^T]$, consider the same problem as the preceding, but let the gyro drift rate plus bias be put through a first order lag $[b/(s + b)]$ before becoming available as measured output. Figure 5.21 presents a state model block diagram in untransformed coordinates. The augmented system equations are:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} -\alpha & 0 & 0 \\ 0 & 0 & 0 \\ b & b & -b \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} w(t)$$

$$z(t) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} x(t)$$

Since the white noise enters the model more than one integration back from the measurement, $[HGQG^TH^T]$ is singular:

$$HGQG^TH^T = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} Q \begin{bmatrix} \alpha & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 0$$

To circumvent this singularity problem, define a "new measurement"

$$z'(t) = (1/b)\dot{z}(t) + z(t)$$
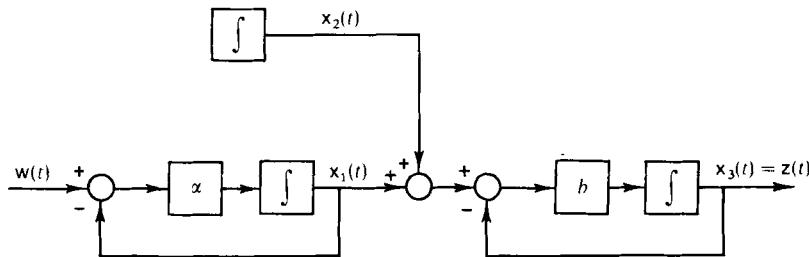
FIG. 5.21   System model for Example 5.15.

Because of the differentiation of $z(t)$ in this definition, the new measurement is separated by only one integration from the white noise. In fact, in the Laplace domain, $z'(s) = [(s + b)/b]z(s)$, or

$$z'(t) = x_1(t) + x_2(t)$$

We have already processed this measurement in the previous example.

Note that if a white noise entered the model at the same location where $x_2(t)$ enters, the Deyst filter could be employed with no difficulty. Moreover, if the output of the lag $[b/(s + b)]$ were corrupted by white noise, a Kalman filter could be used directly.   ∎

5.12   WIENER FILTERING
   AND FREQUENCY DOMAIN TECHNIQUES

The purpose of this section is to relate the optimal filtering results obtained by time-domain (state space) methods to frequency domain techniques. Furthermore, the pioneering work of Wiener [34, 36, 83] will be presented, culminating in the Wiener–Hopf equation. His original problem formulation and the Wiener–Hopf equation itself are in the time domain, but under restrictions (namely, system time-invariance, noise stationarity, and infinite data length) that allow frequency domain interpretation. Although these results have been extended to less restrictive assumptions, the most useful means of solving the Wiener–Hopf equation to yield systematic design capability, the Bode–Shannon technique, is a frequency domain procedure. In the case of time-invariant system models and stationary noises, the Wiener filter will be shown to be equivalent to the steady state Kalman or Deyst filter appropriate to the given problem. Throughout this section, the power spectrum and frequency domain system concepts developed in Sections 4.3 and 4.12 will be exploited.

First the optimal linear estimation problem will be discussed in the communication theory context and notation appropriate to Wiener filter development. Assume that an input signal $i(\cdot,\cdot)$ is the sum of a wanted signal $s(\cdot,\cdot)$ and an unwanted noise $n(\cdot,\cdot)$

$$i(t) = s(t) + n(t) \qquad\qquad (5\text{-}155)$$

for all $t \in T$. The signal may in fact be deterministic, but can in general be a stochastic process. It is desired to generate the device which will accept $i(t)$

as an input and yield $s(t + \Delta t)$ or some function of it as an output. (If $\Delta t = 0$, the device is called a *filter*; if $\Delta t > 0$ the device is a *predictor*; if $\Delta t < 0$, the device is a *smoother*. We will pursue the concept of a filter.)

Despite this desired function, the device that will actually be generated will accept $i(t)$ and produce an output $y(t)$ for all $t \in T$. The desired output of the device is $d(t)$, the output of some specified linear system described by impulse response function $T_d(\cdot)$ in response to be being driven by $s(t)$. For each sample $s(t, \omega_j) = s(t)$ for all $t \in T$, $d(t)$ is

$$d(t) = \int_{-\infty}^{t} T_d(t - \tau)s(\tau) \, d\tau \qquad (5\text{-}156)$$

(Note that time-invariance is assumed by writing $T_d(\cdot)$ instead of $T_d(\cdot, \cdot)$, but this can be generalized.) In most cases of interest, $T_d(\tau)$ is $\delta(\tau)$ for all τ: in other words, $d(t) = s(t)$, which is to say that we *desire* filter output $y(t)$ to be the signal $s(t)$ itself.

Thus, a block diagram of the optimal filtering problem would be as depicted in Fig. 5.22. Note that we cannot really separate out $s(t)$ to put through the "desired operation," or else there would be no filtering problem at all. The filter error, $e(t)$, is the difference between the actual filter output $y(t)$ and the desired output $d(t)$

$$e(t) = y(t) - d(t) \qquad (5\text{-}157)$$

or, in most cases, $[y(t) - s(t)]$. From the many possible criteria for optimality, that of least mean square error is particularly tractable if one can specify process autocorrelation functions and the system impulse response function or transfer function of the desired operation.

In the general problem formulation, the characterization of the input and noise is important; if these are stochastic processes, it is particularly important to know whether they are stationary or not. Also important is the amount of design freedom: do we assume that the filter structure is to be linear or that it is time-invariant? To ensure practical value of the design, it is usually required that the filter be realizable (nonanticipative): that the optimal filter impulse response function satisfy $T_{FO}(\tau) = 0$ for all $\tau < 0$. (This was inherent in the
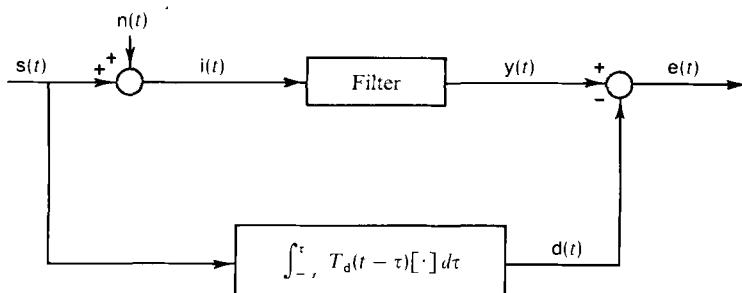
FIG. 5.22   Optimal filtering problem in communication theory context.

recursive state nature of the Kalman filter.) A final design parameter of the filter is its operating time: whether it operates on a finite measurement sample length or if it represents a steady state solution for an infinite length of measurement data.

In 1942, Wiener solved the filtering problem under the following assumptions: (1) the signal $s(t)$ and noise $n(t)$ are each samples from stationary random processes which have some distinguishing statistical characteristics, (2) the filter is a (time-invariant) linear device that operates on an infinite record of data, and (3) the optimality criterion is minimum mean square error. Note that only time-invariant devices need be considered since we seek a filter that produces an output with stationary statistics in response to a stationary input. Denoting a given filter impulse response function by $T_F(\cdot)$, realizations of (5-157) can be written as

$$e(t) = y(t) - d(t) = \int_{-\infty}^{\infty} T_F(\tau)i(t - \tau)\,d\tau - d(t)$$

where the limits of integration are valid because of infinite data record length and realizability. Thus, the mean square error for any filter is

$$E\{e^2\} = E\{y^2\} - 2E\{yd\} + E\{d^2\} = \Psi_{yy}(0) - 2\Psi_{yd}(0) + E\{d^2\}$$

$$= \int_{-\infty}^{\infty} T_F(\tau_1)\left[\int_{-\infty}^{\infty} T_F(\tau_2)\Psi_{ii}(\tau_1 - \tau_2)\,d\tau_2\right]d\tau_1$$

$$- 2\int_{-\infty}^{\infty} T_F(\tau)\Psi_{id}(\tau)d\tau + E\{d^2\} \qquad (5\text{-}158)$$

where $\Psi_{xz}(\tau)$ denotes the correlation function of $x(t)$ and $z(t + \tau)$, $E\{x(t)z(t + \tau)\}$. By employing variational techniques, the filter may be written as the sum of the optimal filter described by impulse response function $T_{FO}(\cdot)$ plus a perturbation described by $\Delta T_F(\cdot)$:

$$T_F(t) = T_{FO}(t) + \varepsilon \Delta T_F(t)$$

for all time t of interest. This is substituted into (5-158), and then the necessary condition for a minimum is

$$0 = \frac{\partial E\{e^2\}}{\partial \varepsilon}\bigg|_{\varepsilon = 0} = 2\int_{-\infty}^{\infty} \Delta T_F(t)\left[\int_{-\infty}^{\infty} T_{FO}(\tau)\Psi_{ii}(t - \tau)d\tau - \Psi_{id}(t)\right]dt$$

Since $\Delta T_F(t)$ is arbitrary for all $t \geq 0$ and zero for $t < 0$, for the above expression to be valid, the term within the brackets must itself equal zero. Thus is obtained the now famous *Wiener–Hopf equation*

$$\int_{-\infty}^{\infty} T_{FO}(\tau)\Psi_{ii}(t - \tau)d\tau - \Psi_{id}(t) = 0 \qquad (5\text{-}159)$$

The solution to this integral equation is not a trivial task. To find $T_{FO}(\cdot)$, given $\Psi_{ii}(\cdot)$ and $\Psi_{id}(\cdot)$, it is usually necessary to employ integral transform techniques and solve for the filter transfer function in the frequency domain.

Perhaps the most useful solution method is the *Bode–Shannon technique* [7], which states that, if the desired transfer function $T_d(s)$ and the signal and noise spectra are all rational, then the solution for the optimal realizable filter is

$$T_{FO}(s) = \frac{1}{\Psi_{ii}(s)_L} \left[\frac{\Psi_{id}(s)}{\Psi_{ii}(s)_R} \right]_{\mathscr{L}} \qquad (5\text{-}160)$$

where the subscripts L and R denote *spectral factorization* and $\mathscr{L}$ denotes *separation by partial fraction expansion* (to be explained further as the entire procedure is specified).

The power spectral density of the input is given by

$$\Psi_{ii}(s) = \Psi_{ss}(s) + \Psi_{sn}(s) + \Psi_{ns}(s) + \Psi_{nn}(s) \qquad (5\text{-}161)$$

where, as before, the subscript i denotes input, s refers to signal, and n pertains to noise. If the noise and signal are uncorrelated, as if often the case, then (5-161) becomes

$$\Psi_{ii}(s) = \Psi_{ss}(s) + \Psi_{nn}(s) \qquad (5\text{-}161')$$

This expression can be partitioned by spectral factorization (see Section 4.12) as

$$\Psi_{ii}(s) = \Psi_{ii}(s)_L \Psi_{ii}(s)_R \qquad (5\text{-}162)$$

where $\Psi_{ii}(s)_L$ has all of its poles and zeros confined to the left half s plane (including half the pole and zero doubles on the imaginary axis), and $\Psi_{ii}(s)_R$ similarly has all of its poles and zeros in the right half plane including the other half of the doubles on the $j\omega$ axis). These factors are used directly in (5-160).

The cross power spectral density between the input and the desired output is given by

$$\Psi_{id}(s) = T_d(s)[\Psi_{ss}(s) + \Psi_{ns}(s)] \qquad (5\text{-}163)$$

If the noise and signal are uncorrelated, this becomes:

$$\Psi_{id}(s) = T_d(s)\Psi_{ss}(s) \qquad (5\text{-}163a)$$

We are especially interested in the case of $T_d(s) = 1$, for which

$$\Psi_{id}(s) = \Psi_{is}(s) = \Psi_{ss}(s) + \Psi_{ns}(s) \qquad (5\text{-}163b)$$

and if the noise and signal are also uncorrelated, then this simplifies to

$$\Psi_{id}(s) = \Psi_{ss}(s) \qquad (5\text{-}163c)$$

To use the Bode-Shannon technique, first the expression $[\Psi_{id}(s)/\Psi_{ii}(s)_R]$ is generated, and then it is separated by partial fraction expansion:

$$\left[\frac{\Psi_{id}(s)}{\Psi_{ii}(s)_R} \right] = \sum_{j=1}^{N} \sum_{k=1}^{\mu_j} \frac{A_{jk}}{(s + a_j)^k} \qquad (5\text{-}164)$$

where N is the number of distinct poles and μ_j is the multiplicity of the jth pole.

This sum can then be separated into the sum of terms corresponding to left half plane poles, denoted as $[\Psi_{id}(s)/\Psi_{ii}(s)_R]_{\mathscr{L}}$, and a sum of terms corresponding to right half plane poles, denoted as $[\Psi_{id}(s)/\Psi_{ii}(s)_R]_{\mathscr{R}}$:

$$[\Psi_{id}(s)/\Psi_{ii}(s)_R] = [\Psi_{id}(s)/\Psi_{ii}(s)_R]_{\mathscr{L}} + [\Psi_{id}(s)/\Psi_{ii}(s)_R]_{\mathscr{R}} \qquad (5\text{-}165)$$

The term $[\Psi_{id}(s)/\Psi_{ii}(s)_R]_{\mathscr{L}}$ is then multiplied by $[1/\Psi_{ii}(s)_L]$ to obtain $T_{FO}(s)$, the optimal filter transfer function.

This method is probably the easiest means of solving the Wiener–Hopf equation, but it is valid only for cases involving stationary signal and noise and infinite length data records. Thus, a systematic design technique is available for generating steady state optimal filters in the frequency domain, but its power and applicability is more restricted than that of the Kalman filter synthesis capability.

EXAMPLE 5.16   Assume that we have a signal and noise which are uncorrelated, with power spectral densities given by

$$\Psi_{ss}(\omega) = A/(a^2 + \omega^2), \qquad \Psi_{nn}(\omega) = \Psi_0$$

i.e., an exponentially time-correlated signal corrupted by white noise. We want to design the optimal filter to accept the sum of these two and output the best representation of the signal alone. The desired output is the signal itself, so $T_d(s) = 1$ and

$$\Psi_{id}(s) = \Psi_{ss}(s) = \frac{A}{a^2 - s^2}$$

The input power spectral density is

$$\Psi_{ii}(s) = \Psi_{ss}(s) + \Psi_{nn}(s)$$

$$= \frac{A}{a^2 - s^2} + \Psi_0 = \Psi_0 \frac{(a^2 + A/\Psi_0) - s^2}{a^2 - s^2}$$

Letting $c^2 = a^2 + (A/\Psi_0)$, spectral factorization of $\Psi_{ii}(s)$ yields

$$\Psi_{ii}(s) = \underbrace{\left[\sqrt{\Psi_0}\frac{c + s}{a + s}\right]}_{\Psi_{ii}(s)_L} \underbrace{\left[\sqrt{\Psi_0}\frac{c - s}{a - s}\right]}_{\Psi_{ii}(s)_R}.$$

Thus, the term $[\Psi_{id}(s)/\Psi_{ii}(s)_R]$ becomes

$$\frac{\Psi_{id}(s)}{\Psi_{ii}(s)_R} = \left[\frac{A}{(a + s)(a - s)}\right]\left[\frac{1}{\sqrt{\Psi_0}}\frac{a - s}{c - s}\right] = \frac{A}{\sqrt{\Psi_0}}\left[\frac{1}{(a + s)(c - s)}\right]$$

This can be factored into the partial fraction expansion:

$$\left[\frac{\Psi_{id}(s)}{\Psi_{ii}(s)_R}\right] = \frac{A}{\sqrt{\Psi_0}}\left[\frac{1/(a + c)}{a + s} + \frac{1/(a + c)}{c - s}\right]$$

so that

$$\left[\frac{\Psi_{id}(s)}{\Psi_{ii}(s)_R}\right]_{\mathscr{L}} = \frac{A}{\sqrt{\Psi_0}(a + c)}\frac{1}{a + s}$$

Multiplying this by $1/\bar{\Psi}_{ii}(s)_L$ yields $T_{FO}(s)$ as

$$T_{FO}(s) = \left[\frac{1}{\sqrt{\Psi_0}}\frac{a+s}{c+s}\right]\left[\frac{A}{\sqrt{\Psi_0}(a+c)}\frac{1}{a+s}\right]$$

$$= \frac{A}{\Psi_0(a+c)}\frac{1}{c+s} = \frac{A(c-a)}{\Psi_0(c^2-a^2)}\frac{1}{c+s}$$

$$= \frac{c-a}{s+c}$$

where $c = \sqrt{a^2 + (A/\bar{\Psi}_0)}$.

Let us investigate the reasonableness of this first order lag as the optimal filter form. The break frequency of the filter is $c = \sqrt{a^2 + (A/\bar{\Psi}_0)}$. If the low frequency signal-to-noise ratio is much greater than one:

$$A/(\bar{\Psi}_0 a^2) \gg 1$$

then the break frequency is approximately

$$c \cong \sqrt{A/\bar{\Psi}_0} \gg a$$

and the magnitude of the filter transfer function is approximately one at low frequencies. Thus for the case in which the input contains a large proportion of valid information, the filter pays considerable attention to the input and does not attenuate it significantly until a frequency considerably beyond the signal break frequency, a.

When the signal-to-noise ratio is small,

$$A/(\bar{\Psi}_0 a^2) \ll 1$$

then $c \cong a$ and the low frequency gain is approximately $\frac{1}{2}[A/(\bar{\Psi}_0 a^2)] \ll 1$. Thus, the filter attenuates the input and breaks where the signal breaks. ∎

Numerous attempts were made to extend Wiener's work by relaxing some of his assumptions [8, 71, 86]. By 1953 the necessary condition for a time-varying optimal filter operating on a finite length data record (of length Δt) generated by samples from nonstationary stochastic process inputs was shown to be

$$\int_0^{\Delta t} T_{FO}(\tau, \sigma)\Psi_{ii}(\sigma - t, \sigma - \tau)\,d\tau - \Psi_{id}(\sigma - t, \sigma) = 0 \qquad (5\text{-}166)$$

Although very general problems had been formulated and necessary conditions for optimality obtained in the form of integral equations, the solution to these equations is extremely difficult (if at all tractable) for all but the original, more restricted, Wiener–Hopf equation, (5-159). It would not be until 1960, with the advent of state space time–domain methods, the Kalman filter, and digital computers, that a practical design procedure would be available to generate optimal filters capable of operating on finite data samples of nonstationary stochastic process inputs.

The Kalman filter can readily be applied in this communication theory context. The wanted signal $\mathbf{s}(t)$ is considered to be $\mathbf{H}(t)\mathbf{x}(t)$, the output of a linear system driven by white noise: this is justifiable if the signal spectrum is assumed to be (well approximated as) *rational*. The corruptive noise which is

added to the signal is also considered to be *white* noise, $\mathbf{v}_c(\cdot, \cdot)$. Actually, by means of shaping filters, this formulation can be generalized to allow the corruptive noise to be the *sum* of time-correlated and white noises (if the corruption is composed only of time-correlated noise, then a Deyst filter must be used instead).

The continuous-time Kalman filter is given by (5-144)–(5-146) without the deterministic noise term, and in general is a time-varying system. If the system model is time invariant ($\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ constant) and the noises are stationary ($\mathbf{Q}$ and $\mathbf{R}$ constant), the filter may reach steady state performance in which the covariance $\mathbf{P}$ is a constant (sufficient conditions for stability being given in Section 5.11). For this condition of $\dot{\mathbf{P}}(t) = \mathbf{0}$, the Riccati equation becomes an algebraic relation

$$\dot{\mathbf{P}} = \mathbf{FP} + \mathbf{PF}^T + \mathbf{GQG}^T - \mathbf{PH}^T\mathbf{R}_c^{-1}\mathbf{HP} = \mathbf{0} \qquad (5\text{-}167)$$

In the steady state condition, the rate at which uncertainty increases (given by $\mathbf{GQG}^T$) is just balanced by the rate at which new information enters ($\mathbf{PH}^T\mathbf{R}_c^{-1}\mathbf{HP}$) and the dissipative effects of the system ($\mathbf{FP} + \mathbf{PF}^T$). For a steady state covariance matrix, the optimal filter is also time invariant, given by

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{F}\hat{\mathbf{x}}(t) + \mathbf{PH}^T\mathbf{R}_c^{-1}[\mathbf{z}(t) - \mathbf{H}\hat{\mathbf{x}}(t)] \qquad (5\text{-}168\text{a})$$

$$= [\mathbf{F} - \mathbf{KH}]\hat{\mathbf{x}}(t) + \mathbf{Kz}(t) \qquad (5\text{-}168\text{b})$$

Taking the Laplace transform of this (neglecting initial conditions) yields

$$(s\mathbf{I} - \mathbf{F} + \mathbf{KH})\hat{\mathbf{x}}(s) = \mathbf{Kz}(s)$$

so that

$$\hat{\mathbf{x}}(s) = [(s\mathbf{I} - \mathbf{F} + \mathbf{KH})^{-1}\mathbf{K}]\mathbf{z}(s) \qquad (5\text{-}169)$$

where the term in brackets is the transfer function representation of the steady state Kalman filter [56]. It is identical to the Wiener filter found for the case of white corruptive noise by solving the Wiener–Hopf equation (5-159) by the Bode–Shannon technique. In order words, *the steady state Kalman filter is equivalent to the Wiener filter for the same problem formulation.*

EXAMPLE 5.17   Consider the same problem as previously solved by Wiener filter design in Example 5.16, be now determine the Kalman filter to generate an optimal estimate of the wanted signal. First it is necessary to determine the shaping filter which will generate a stochastic process with the given power spectral density $\Psi_{xx}(\omega) = A/(a^2 + \omega^2)$. This is found to be a first order lag described by

$$G(s) = 1/(s + a)$$

driven by white noise with mean zero and strength A.

Thus, the system model upon which the filter is based is

$$\dot{\mathbf{x}}(t) = -a\mathbf{x}(t) + \mathbf{w}(t), \qquad \mathbf{z}_c(t) = \mathbf{x}(t) + \mathbf{v}_c(t)$$

with

$$E\{w(t)w(t+\tau)\} = A\,\delta(\tau), \qquad E\{v_c(t)v_c(t+\tau)\} = \Psi_0\,\delta(\tau) \qquad E\{w(t)v_c(t+\tau)\} = 0$$

The Kalman filter is specified by

$$\dot{\hat{x}}(t) = -a\hat{x}(t) + [P(t)/\Psi_0][z(t) - \hat{x}(t)]$$

where $P(t)$ satisfies the Riccati equation

$$\dot{P}(t) = -2aP(t) + A - [P^2(t)/\Psi_0]$$

In steady state, the variance is a constant, P, given by

$$P = \Psi_0[\sqrt{a^2 + (A/\Psi_0)} - a] = \Psi_0[c - a]$$

Thus, the steady state filter becomes a time-invariant system described by the differential equation:

$$\begin{aligned}
\dot{\hat{x}}(t) &= -[a + (P/\Psi_0)]\hat{x}(t) + [P/\Psi_0]z(t)\\
&= -[a + c - a]\hat{x}(t) + [c - a]z(t)\\
&= -[c]\hat{x}(t) + [c - a]z(t)
\end{aligned}$$

or equivalently by the Laplace transfer function

$$\frac{\hat{x}(s)}{z(s)} = \frac{c-a}{s+c}$$

This is identical to the result of Example 5.16.

Note again that this is a *steady state* Kalman filter. $P(t)$ is a complicated time function:

$$P(t) = \Psi_0\sqrt{a^2 + (A/\Psi_0)}\tanh(t+k) - \Psi_0 A$$
$$k = \tanh^{-1}\{[a + (P_0/\Psi_0)]/\sqrt{a^2 + (A/\Psi_0)}\}$$

and the filter is generally a time-varying linear system.   ∎

Furthermore, *the steady state Deyst filter is equivalent to the Wiener filter for the same problem formulation.* For a system model given by (5-137) and (5-152), the Deyst filter was given by (5-153) and (5-154). For the particular case of a time-invariant system driven by stationary noise $w(\cdot, \cdot)$, a steady state solution can be reached:

$$\hat{x}(t) = y(t) + Wz(t) \tag{5-170a}$$

$$\dot{y}(t) = [I - WH]F\hat{x}(t) \tag{5-170b}$$

$$A = [PF^T + GQG^T]H^T \tag{5-170c}$$

$$W = A[HGQG^TH^T]^{-1} \tag{5-170d}$$

$$0 = FP + PF^T + GQG^T - WA^T \tag{5-170e}$$

The Laplace transform transfer function of this filter is readily shown to be

$$\hat{x}(s) = \{[sI - (I - WH)F]^{-1}[I - WH]FW + W\}z(s) \tag{5-171}$$

This can be shown to be identical to the Wiener filter for the case of time-correlated measurement noise only. Note the direct feedthrough term W in

(5-171): unlike the Kalman filter, the steady state Deyst filter transfer function matrix can have numerators of order equal to the corresponding denominators, as previously claimed. The equivalency of the two forms is demonstrated for a particular application in Problems 5.26 and 5.27.

5.13 SUMMARY

This chapter formulated and solved the optimal estimation problem for the case in which a linear system model driven by white Gaussian noises and deterministic inputs adequately describes true system behavior. Because of its practical applicability through digital computer implementation, the sampled data formulation was emphasized throughout. Table 5.2 summarizes the Kalman filter algorithm for discrete-time measurements, comprised of relations for propagating the state estimate and error covariance from one measurement time to the next, and update equations for incorporating the next measurement

TABLE 5.2

Kalman Filter for Discrete-Time Measurements

State Dynamics Model	Time Propagation Relations
(1) Stochastic differential equation	
$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\mathbf{w}(t)$	$\dot{\hat{\mathbf{x}}}(t/t_{i-1}) = \mathbf{F}(t)\hat{\mathbf{x}}(t/t_{i-1}) + \mathbf{B}(t)\mathbf{u}(t)$
$\mathbf{dx}(t) = \mathbf{F}(t)\mathbf{x}(t)\,dt + \mathbf{B}(t)\mathbf{u}(t)\,dt$	$\dot{\mathbf{P}}(t/t_{i-1}) = \mathbf{F}(t)\mathbf{P}(t/t_{i-1}) + \mathbf{P}(t/t_{i-1})\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t)$
$\qquad + \mathbf{G}(t)\,\mathbf{d\beta}(t)$	
(2) Solution to stochastic differential equation	
$\mathbf{x}(t) = \mathbf{\Phi}(t,t_0)\mathbf{x}(t_0)$	$\hat{\mathbf{x}}(t_i^-) = \mathbf{\Phi}(t_i,t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+) + \int_{t_{i-1}}^{t_i} \mathbf{\Phi}(t_i,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$
$\qquad + \int_{t_0}^{t} \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$	$\mathbf{P}(t_i^-) = \mathbf{\Phi}(t_i,t_{i-1})\mathbf{P}(t_{i-1}^+)\mathbf{\Phi}^{\mathrm{T}}(t_i,t_{i-1})$
$\qquad + \int_{t_0}^{t} \mathbf{\Phi}(t,\tau)\mathbf{G}(\tau)\,\mathbf{d\beta}(\tau)$	$\qquad + \int_{t_{i-1}}^{t_i} \mathbf{\Phi}(t_i,\tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\mathbf{\Phi}^{\mathrm{T}}(t_i,\tau)\,d\tau$
(3) Stochastic difference equation	
$\mathbf{x}(t_i) = \mathbf{\Phi}(t_i,t_{i-1})\mathbf{x}(t_{i-1})$	$\hat{\mathbf{x}}(t_i^-) = \mathbf{\Phi}(t_i,t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+) + \mathbf{B}_\mathrm{d}(t_{i-1})\mathbf{u}(t_{i-1})$
$\qquad + \mathbf{B}_\mathrm{d}(t_{i-1})\mathbf{u}(t_{i-1})$	$\mathbf{P}(t_i^-) = \mathbf{\Phi}(t_i,t_{i-1})\mathbf{P}(t_{i-1}^+)\mathbf{\Phi}^{\mathrm{T}}(t_i,t_{i-1})$
$\qquad + \mathbf{G}_\mathrm{d}(t_{i-1})\mathbf{w}_\mathrm{d}(t_{i-1})$	$\qquad + \mathbf{G}_\mathrm{d}(t_{i-1})\mathbf{Q}_\mathrm{d}(t_{i-1})\mathbf{G}_\mathrm{d}^{\mathrm{T}}(t_{i-1})$

Measurement Model	Measurement Update Equations
$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i)$	$\mathbf{K}(t_i) = \mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i)]^{-1}$
	$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$
	$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i^-)$

into the estimate. Three forms of time propagation relations are enumerated in the table: (1) differential equations to be solved numerically, based on the stochastic differential equation description of state dynamics, (2) the discrete-time algorithm based upon the solution to the original stochastic differential equation, and (3) the discrete-time algorithm based upon a stochastic difference equation (especially viewed as an equivalent discrete-time model to the original continuous-time dynamics). By recursively generating the conditional mean and error covariance, the Kalman filter actually maintains an explicit description of the entire conditional density for the states conditioned on the entire measurement history, thereby fulfilling the objective of Bayesian estimation. The analytical and computational characteristics of the Kalman filter were analyzed, alternate and extended forms explored, continuous-time measurements considered and equivalency to Wiener filtering in the steady state investigated. The next chapter will further explore the design and implementation of practical Kalman filters.

REFERENCES

1. Albert, A., and Sittler, R. W., "A Method for Computing Least Squares Estimators that Keep Up with the Data," *SIAM J. Control* 3, 384–417 (1966).
2. Aoki, M., and Huddle, J. R., "Estimation of the State Vector of a Linear Stochastic System with a Constrained Estimator," *IEEE Trans. Automatic Control* **AC-12**, 432–433 (1967).
3. Athans, M., and Tse, E., "A Direct Derivation of the Optimal Linear Filter Using the Maximum Principle," *IEEE Trans. Automatic Control* **AC-12**, 690–698 (1967).
4. Battin, R. H., "A Statistical Optimizing Navigation Procedure for Space Flight," *J. Amer. Rocket Soc.* **32**, 1681–1692 (1962).
5. Battin, R. H., *Astronautical Guidance*, pp. 303–340. McGraw-Hill, New York, 1964.
6. Beavers, A. N., and Denman, E. D., "A New Solution Method for Quadratic Matrix Equations," *Math. Biosci.* **20**, 135–143 (1974).
7. Bode, H. W., and Shannon, C. E., "A Simplified Derivation of Linear Least-Squares Smoothing and Prediction Theory," *Proc. IRE* **38**, 417–425 (1950).
8. Booton, R. C., "On Optimization Theory for Time-Varying Linear Systems with Nonstationary Statistical Inputs," *Proc. IRE* **40**, 977–981 (1952); also M.I.T. Meteor Rep. 72 (July 1951).
9. Box, G. E. P., and Jenkins, G. M., *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, California, 1976.
10. Bryson, A. E. Jr., and Henrikson, L. J., "Estimation Using Sampled-Data Containing Sequentially Correlated Noise," Tech. Rep. No. 533, Division of Engineering and Applied Physics, Harvard Univ., Cambridge, Massachusetts, June 1967.
11. Bryson, A. E. Jr., and Johansen, D. E., "Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise," *IEEE Trans. Automatic Control* **AC-10**, 4–10 (1965).
12. Bucy, R. S., "Global Theory of the Riccati Equation," *J. Comput. System Sci.* **1**, 349–361 (1967).
13. Bucy, R. S., "Optimal Filtering for Correlated Noise," *J. Math. Anal. Appl.* **20**, 1–8 (1967).
14. Cox, H., "On the Estimation of State Variables and Parameters for Noisy Dynamic Systems," *IEEE Trans. Automatic Control* **AC-9**, 5–12 (1964).
15. Cox, H., "Estimation of State Variables Via Dynamic Programming," *Proc. Joint Automatic Control Conf., Stanford, California*, 376–381 (1964).

16. Cox, H., "Sequential Minimax Estimation," *IEEE Trans. Automatic Control* **AC-11**, 323–324 (1966).

17. Cramér, H., *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, New Jersey, 1966.

18. Demetry, J. S., "A Note on the Nature of Optimality in the Discrete Kalman Filter," *IEEE Trans. Automatic Control* **AC-15**, 603–604 (1970).

19. Denman, E. D., and Beavers, A. N., "The Matrix Sign Function and Computations in Systems," *Appl. Math. Comput.* **2**, 63–94 (1976).

20. Deyst, J. J., "Optimum Continuous Estimation of Nonstationary Random Variables," Rep. T-369. M.I.T. Instrumentation Laboratory, Cambridge, Massachusetts, January 1964.

21. Deyst, J. J., "A Derivation of the Optimum Continuous Linear Estimator for Systems with Correlated Measurement Noise," *AIAA J.* **7**, 2116–2119 (1969).

22. Deyst, J. J., and Price, C. F., "Conditions for the Asymptotic Stability of the Discrete, Minimum Variance, Linear Estimator," *IEEE Trans. Automatic Control* **AC-13**, 702–705 (1968).

23. Doob, J. L., *Stochastic Processes*. Wiley, New York, 1953.

24. Fagin, S. L., "Recursive Linear Regression Theory, Optimal Filter Theory, and Error Analysis of Optimal Systems," *IEEE Internat. Convention Record, New York*, 216–240 (1964).

25. Fitzgerald, R. J., "Divergence of the Kalman Filter," *IEEE Trans. Automatic Control* **AC-16**, 736–747 (1971).

26. Friedland, B., "Treatment of Bias in Recursive Filtering," *IEEE Trans. Automatic Control* **AC-14**, 359–367 (1969).

27. Gauss, K. F., "Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections," (1869), reprinted by Dover, New York, 1963.

28. Ho, Y., "The Method of Least Squares and Optimal Filtering Theory," The Rand Corporation, Memorandum RM-3329-PR, Santa Monica, California, October 1962.

29. Ho, Y. C., "On the Stochastic Approximation Method and Optimal Filtering Theory," *J. Math. Anal. Appl.* **6**, 152–154 (1963).

30. Ho, Y. C., and Lee, R. C. K., "A Bayesian Approach to Problems in Stochastic Estimation and Control," *Proc. Joint Automatic Control Conf., Stanford, California*, 382–387 (1964).

31. Jazwinski, A. H., *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

32. Kailath, T., "An Innovations Approach to Least Squares Estimation—Part I: Linear Filtering in Additive White Noise," *IEEE Trans. Automatic Control* **AC-13**, 646–655 (1968).

33. Kailath, T., "The Innovations Approach to Detection and Estimation Theory," *Proc. IEEE* **58**, 680–695 (1970).

34. Kailath, T., and Geesey, R. A., "An Innovations Approach to Least Squares Estimation— Part IV: Recursive Estimation Given Lumped Covariance Functions," *IEEE Trans. Automatic Control* **AC-16**, 720–726 (1971).

35. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," *Trans. ASME J. Basic Eng.* **82**, 34–45 (1960).

36. Kalman, R. E., "New Methods in Wiener Filtering Theory," *Proc. Symp. Eng. Appl. of Random Function Theory and Probability, 1st*. Wiley, New York, 1963.

37. Kalman, R. E., and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," *Trans. ASME J. Basic Eng.* **83**, 95–108 (1961).

38. Kleinman, D. L., "On the Linear Regulator Problem and the Matrix Riccati Equation," Tech. Rep. ESL-R-271. M.I.T. Electronic Systems Laboratory, Cambridge, Massachusetts, June 1966.

39. Kleinman, D. L., "On an Iterative Technique for Riccati Equation Computations," *IEEE Trans. Automatic Control* **AC-13**, 114–115 (1968).

40. Kleinman, D. L., "Iterative Solution of Algebraic Riccati Equations," *IEEE Trans. Automatic Control* **AC-19**, 252–254 (1974).

41. Kleinman, D. L., "A Description of Computer Programs Useful in Linear Systems Studies," Tech. Rep. TR-75-4. University of Connecticut, Storrs, Connecticut, October 1975.

42. Kwakernaak, H., and Sivan, R., *Linear Optimal Control Systems*. Wiley, New York, 1972.

43. Lainiotis, D. G., "Discrete Riccati Equation Solutions: Partitioned Algorithms," *IEEE Trans. Automatic Control* **AC-20**, 555–556 (1975).

44. Lainiotis, D. G., "Partitioned Riccati Solutions and Integration-Free Doubling Algorithms," *IEEE Trans. Automatic Control* **AC-21**, 677–689 (1976).

45. Luenberger, D. G., "Observing the State of a Linear System," *IEEE Trans. Military Electron.* **MIL-8**, 74–80 (1964).

46. Luenberger, D. G., "Observers for Multivariable Systems," *IEEE Trans. Automatic Control* **AC-11**, 190–197 (1966).

47. Luenberger, D., *Optimization by Vector Space Methods*. Wiley, New York, 1969.

48. Luenberger, D. G., "An Introduction to Observers," *IEEE Trans. Automatic Control* **AC-16**, 596–602, (1971).

49. Maybeck, P. S., "Combined Estimation of States and Parameters for On-Line Applications," Ph.D. dissertation, M.I.T., Cambridge, Massachusetts, February 1972.

50. Maybeck, P. S., "The Kalman Filter—An Introduction for Potential Users," TM-72-3, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, June 1972.

51. Maybeck, P. S., "Failure Detection Through Functional Redundancy," Technical Rep. AFFDL-TR-74-3, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, January 1974.

52. Maybeck, P. S., "Failure Detection Without Excessive Hardware Redundancy," *Proc. IEEE Nat. Aerospace and Electron. Conf., Dayton, Ohio* (May 1976).

53. McGarty, T. P., *Stochastic Systems and State Estimation*. Wiley, New York, 1974.

54. Meditch, J. S., *Stochastic Optimal Linear Estimation and Control*. McGraw-Hill, New York, 1969.

55. Mehra, R. K., and Peschon, J., "An Innovations Approach to Fault Detection and Diagnosis in Dynamic Systems," *Automatica* **7**, 637–640 (1971).

56. Melsa, J. L., "Frequency Domain Derivation of the Stationary Kalman Filter," *Proc. Annu. Houston Conf. Circuits, Systems, and Comput., 1st* pp. 323–332 (1969).

57. Mowery, V. O., "Least Squares Recursive Differential-Correction Estimation in Nonlinear Problems," *IEEE Trans. Automatic Control* **AC-10**, 399–407 (1965).

58. Potter, J. E., "A Matrix Equation Arising in Statistical Filter Theory," Rep. RE-9. M.I.T. Experimental Astronomy Laboratory, Cambridge, Massachusetts, February 1965.

59. Potter, J. E., "Matrix Quadratic Solutions," *SIAM J. Appl. Math.* **14**, 496–501 (1966).

60. Potter, J. E., and Stern, R. G., "Statistical Filtering of Space Navigation Measurements," *Guidance and Control—II* (R. C. Langford and C. J. Mundo eds.), Vol 13 of Progress in in Aeronautics and Astronautics, Academic Press, New York, 1964.

61. Potter, J. E., and Vander Velde, W. E., "Optimum Mixing of Gyroscope and Star Tracking Data," Rep. RE-26. M.I.T. Experimental Astronomy Laboratory, Cambridge, Massachusetts; also *J. Spacecraft and Rockets* **5**, 536–540 (1968).

62. Rauch, H. E., Tung, F., and Striebel, C. T., "Maximum Likelihood Estimates of Linear Dynamical Systems," *AIAA J.* **3**, 1445–1450 (1965).

63. Repperger, D. W., "A Square Root of a Matrix Approach to Obtain the Solution to a Steady State Matrix Riccati Equation," *IEEE Trans. Automatic Control* **AC-21**, 786–787 (1976).

64. Rhodes, I. B., "A Tutorial Introduction to Estimation and Filtering," *IEEE Trans. Automatic Control* **AC-16**, 688–706 (1971).

65. Sage, A. P., and Masters, G. W., "Least Squares Curve Fitting and Discrete Optimal Filtering," *IEEE Trans. Education* **E-10**, 29–36 (1967).

66. Schmidt, S. F., "Applications of State-Space Methods to Navigation Problems," *Advances in Control Systems*, Vol. 3. Academic Press, New York, 1966.

67. Schweppe, F. C., "Evaluation of Likelihood Functions for Gaussian Signals," *IEEE Trans. Information Theory* IT-11, 61–70 (1965).

68. Schweppe, F. C., "Recursive State Estimation: Unknown but Bounded Errors and System Inputs," *IEEE Trans. Automatic Control* AC-13, 22–28 (1968).

69. Sherman, S., "A Theorem on Convex Sets with Applications," *Ann. Math. Statist.* 26, 763–767 (1955).

70. Sherman, S., "Non-Mean-Square Error Criteria," *IRE Trans. Information Theory* IT-4, 125–126 (1958).

71. Shinbrot, M., "Optimization of Time-Varying Linear Systems with Nonstationary Inputs," *Trans. ASME* 80, 457–462 (1958).

72. Sorenson, H. W., "Kalman Filtering Techniques," *Advances in Control Systems*, Vol. 3, pp. 219–292. Academic Press, New York, 1966.

73. Sorenson, H. W., "Least-Squares Estimation: From Gauss to Kalman," *IEEE Spectrum* 63–68 (1970).

74. Stear, E. B., and Stubberud, A. R., "Optimal Filtering for Gauss–Markov Noise," *Internat. J. Control* 8, 123–130 (1968).

75. Swerling, P., "First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations," *J. Astronaut. Sci.* 6, 46–52 (1959).

76. Swerling, P., "Topics in Generalized Least Squares Signal Estimation," *SIAM J. Appl. Math.* 14, 998–1031 (1966).

77. Swerling, P., "Modern State Estimation Methods from the Viewpoint of the Method of Least Squares," *IEEE Trans. Automatic Control* AC-16, 707–719 (1971).

78. Tse, E., "On the Optimal Control of Linear Systems with Incomplete Information," Rep. ESL-R-412. M.I.T. Electronic Systems Laboratory, Cambridge, Massachusetts, 1970.

79. Tse, E., and Athans, M., "Optimal Minimal-Order Observer-Estimators for Discrete Linear Time-Varying Systems," *IEEE Trans. Automatic Control* AC-15, 416–426 (1970).

80. Tse, E., and Athans, M., "Observer Theory for Continuous-Time Linear Systems," *Information and Control* 22, 405–434 (1973).

81. Uttam, B., "Observer Theory," Tech. Rep. EM-163. The Analytic Sciences Corporation, Reading, Massachusetts, April 1970.

82. Van Trees, H. L., *Detection, Estimation and Modulation Theory*, Vol. 1. Wiley, New York, 1969.

83. Wiener, N., *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, OSRD 370, Report to the Services 19, Research Project DIC-6037. M.I.T., Cambridge, Massachusetts, February 1942, also Wiley, New York, 1949.

84. Wilks, S. S., *Mathematical Statistics*. Wiley, New York, 1962.

85. Womble, M. E., and Potter, J. E., "A Prefiltering Version of the Kalman Filter with New Numerical Integration Formulas for Riccati Equations," *Proc. IEEE Conf. Decision and Control*, San Diego, California, pp. 63–67 (December 1973).

86. Zadeh, L. A., and Ragazzini, J. R., "An Extension of Wiener's Theory of Prediction," *J. Appl. Phys.* 21, 645–655 (1950).

PROBLEMS

5.1 The two-dimensional random vector $\mathbf{x} = [x_1 x_2]^T$ has the probability density:

$$f_\mathbf{x}(\boldsymbol{\xi}) = \frac{1}{2\pi|\mathbf{P}|^{1/2}} \exp\{-\tfrac{1}{2}\boldsymbol{\xi}^T\mathbf{P}^{-1}\boldsymbol{\xi}\}; \qquad \mathbf{P} = \begin{bmatrix} 1 & \tfrac{1}{2} \\ \tfrac{1}{2} & 1 \end{bmatrix}$$

A perfect measurement of x_1 is obtained as $z(\omega_j) = x_1(\omega_j) = z = 1$. What is the probability density of the vector $\mathbf{x}$, conditioned on the measurement $z(\omega_j) = z$?

5.2   Prove the matrix inversion lemma, (5-28): for $\mathbf{P}$ and $\mathbf{R}$ positive definite,

$$(\mathbf{P}^{-1} + \mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H})^{-1} = \mathbf{P} - \mathbf{P}\mathbf{H}^{\mathsf{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathsf{T}} + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}$$

Show this by looking at the partitioned matrix

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{P}^{-1} & \mathbf{H}^{\mathsf{T}} \\ \hline \mathbf{H} & -\mathbf{R} \end{array}\right]$$

and letting $\mathbf{A}^{-1}$ be given by the partitioned matrix

$$\mathbf{A}^{-1} = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{F} \\ \hline \mathbf{G}^{\mathsf{T}} & \mathbf{E} \end{array}\right]$$

and solve the equations that result by setting $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ for the value of $\mathbf{D}$ (the upper left partition of $\mathbf{A}^{-1}$).

5.3   Having proven (5-28), use it to establish (5-29) and (5-30).

5.4   In reducing (5-27) to (5-33), it is necessary to show that

$$\frac{|\mathbf{H}\mathbf{P}^{-}\mathbf{H}^{\mathsf{T}} + \mathbf{R}|^{1/2}}{|\mathbf{P}^{-}|^{1/2}|\mathbf{R}|^{1/2}} = \frac{1}{|\mathbf{P}^{+}|^{1/2}}$$

To show this, we exploit three basic properties of determinants:

(1)   if $\mathbf{A}$ and $\mathbf{B}$ are n-by-n, then $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$,

(2) $|\mathbf{A}| = |\mathbf{A}^{\mathsf{T}}|$,

(3)

$$\left|\left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{0} & \mathbf{A}_3 \end{array}\right]\right| = |\mathbf{A}_1||\mathbf{A}_3|$$

(a)   Show that

$$\mathbf{P}^* = \left[\begin{array}{c|c} \mathbf{P}^{-} & \mathbf{P}^{-}\mathbf{H}^{\mathsf{T}} \\ \hline \mathbf{H}\mathbf{P}^{-} & \mathbf{H}\mathbf{P}^{-}\mathbf{H}^{\mathsf{T}} + \mathbf{R} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{P}^{+} & \mathbf{P}^{-}\mathbf{H}^{\mathsf{T}} \\ \hline \mathbf{0} & \mathbf{H}\mathbf{P}^{-}\mathbf{H}^{\mathsf{T}} + \mathbf{R} \end{array}\right]\left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{K}^{\mathsf{T}} & \mathbf{I} \end{array}\right]$$

so that $|\mathbf{P}^*| = |\mathbf{P}^{+}||\mathbf{H}\mathbf{P}^{-}\mathbf{H}^{\mathsf{T}} + \mathbf{R}|$.

(b)   Show that the following matrix inversion of a partitioned, symmetric, positive definite matrix is valid

$$\left[\begin{array}{c|c} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \hline \mathbf{X}_{12}^{\mathsf{T}} & \mathbf{X}_{22} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \mathbf{I} & -\mathbf{X}_{11}^{-1}\mathbf{X}_{12} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right]\left[\begin{array}{c|c} \mathbf{X}_{11}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & (\mathbf{X}_{22} - \mathbf{X}_{12}^{\mathsf{T}}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1} \end{array}\right]\left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\mathbf{X}_{12}^{\mathsf{T}}\mathbf{X}_{11}^{-1} & \mathbf{I} \end{array}\right]$$

and use this to demonstrate that $|\mathbf{P}^*| = |\mathbf{P}^{-}||\mathbf{R}|$.

(c)   Combine (a) and (b) to establish the result.
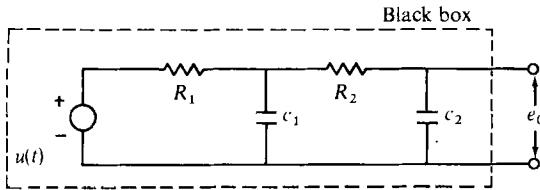
5.5   Verify the numerical results in Table 5.1 of Example 5.4.

5.6   Show that $\mathbf{x}(t_i)$ and $\hat{\mathbf{x}}(t_i^{+})$ are jointly Gaussian, as claimed in Section 5.4. Also show that

$$E\{\mathbf{x}(t_i)\hat{\mathbf{x}}^{\mathsf{T}}(t_i^{+})|\mathbf{Z}(t_i) = \mathbf{Z}_i\} = E\{\mathbf{x}(t_i)|\mathbf{Z}(t_i) = \mathbf{Z}_i\}\hat{\mathbf{x}}^{\mathsf{T}}(t_i^{+}) = \hat{\mathbf{x}}(t_i^{+})\hat{\mathbf{x}}^{\mathsf{T}}(t_i^{+})$$

as claimed in Eq. (5-55).

5.7   Section 5.5 discussed logical definitions of an "optimal" state estimate other than that used to derive the filter results in this chapter. Under what conditions do some of these alternative approaches yield the same value for the "optimal" estimate for a general estimation problem (i.e., no linear model or Gaussian noise assumptions)?

FIG. 5.P1 Circuit for Problem 5.8. $R_1 = R_2 = 1 \, \Omega$; $c_1 = c_2 = 1 \, F$.

5.8 Consider the circuit shown in Fig. 5.P1. It has been constructed and sealed into the proverbial black box. Capacitor c_1 has a very low voltage rating and it is desired to monitor the voltage across c_1 to determine when it exceeds the capacitor limit. The only measurement that can be made on this system is the output voltage, e_0. However, thanks to an exceedingly good voltmeter, essentially perfect measurements can be made of this voltage at discrete times. In order to estimate the voltage across c_1, assume that $u(t)$ can be described as

$$E\{u(t)\} = 0, \qquad E\{u(t_1)u(t_2)\} = Q\,\delta(t_2 - t_1), \qquad Q = 2 \text{ V}^2 \sec$$

Determine an expression for the optimal estimate of the voltage across c_1. Assume that the system starts up with no charge in the capacitors. Plot the variance of the error in the estimate as a function of time, taking measurements every half second for two seconds.

Repeat the solution, assuming the voltmeter output to be the true voltage e_0 plus a zero-mean white Gaussian noise $v(\cdot,\cdot)$ with $E\{v(t_i)v(t_j)\} = R\,\delta_{ij}$, $R = 0.2 \text{ V}^2$.

5.9 Suppose the scalar process $y(\cdot,\cdot)$ satisfied the differential equation

$$\ddot{y}(t) + y(t) = 0$$

where $y(0)$ and $\dot{y}(0)$ are modeled as jointly Gaussian random variables with

$$E[y(0)] = 0, \qquad E[\dot{y}(0)] = 0$$

$$E[y(0)^2] = 4, \quad . \; E[\dot{y}(0)^2] = 2, \qquad E[y(0)\dot{y}(0)] = 1$$

A discrete-time measurement process $z(\cdot,\cdot)$ is available as the output from the system, with

$$z(t_i) = y(t_i) + v(t_i)$$

where $v(\cdot,\cdot)$ is a white Gaussian sequence, independent of $y(0)$ and $\dot{y}(0)$, with

$$E[v(t_i)] = 0, \qquad E[v(t_i)^2] = 1$$

Completely determine the optimal discrete-time estimator for $\dot{y}(t_i)$. What does "optimal" mean here?

Use the difference equation for the error covariance matrix (or the inverse of that matrix) to show that 2π sec is a poor choice of sample period.

5.10 Consider the scalar system model

$$x(t_{i+1}) = x(t_i) + w_d(t_i)$$

where $x(\cdot,\cdot)$ is the state and $w_d(\cdot,\cdot)$ is a discrete-time Gaussian noise with

$$E[w_d(t_i)] = 0, \qquad E[w_d^2(t_i)] = \tfrac{1}{2}, \qquad E[w_d(t_i)w_d(t_j)] = 0 \qquad (i \neq j)$$

The initial state is modeled as Gaussian with statistics

$$E[x(t_1)] = 1, \qquad E[x^2(t_1)] = 2$$

Scalar measurements are available at times t_1 and t_2 as

$$z(t_i) = x(t_i) + v(t_i) \qquad (i = 1, 2)$$

where $v(\cdot, \cdot)$ is a Gaussian sequence with

$$E[v(t_i)] = 0, \qquad E[v^2(t_i)] = \tfrac{1}{4}, \qquad E[v(t_i)v(t_j)] = 0 \qquad (i \neq j)$$

Determine the explicit equations for the optimal estimate of x at time t_1 and t_2: let $z(t_1, \omega_j) = z_1$, and $z(t_2, \omega_j) = z_2$, and obtain explicit equations for the estimates $\hat{x}(t_i^-)$ and $\hat{x}(t_i^+)$, the estimate error variances $P(t_i^-)$ and $P(t_i^+)$, and optimal gain $K(t_i)$ for times t_1 and t_2. What is the value of the "information" added to $P^{-1}(t_i^-)$ by the measurements at times t_1 and t_2?

5.11   Suppose you have a system described by the relation

$$x(t_i) = 0.7x(t_{i-1}) + w_d(t_{i-1}), \qquad t_i = 1, 2, 3 \ldots$$

starting from some known value $x(t_0 = 0) = x_0$, where the $w_d(\cdot, \cdot)$ is a white Gaussian sequence described by statistics

$$E\{w_d(t_i)\} = b = 0.2, \qquad E\{[w_d(t_i) - b]^2\} = 0.01$$

Further suppose that at each sample time t_i, two separate measurements are available:

$$z_1(t_i) = 2x(t_i) + v_1(t_i), \qquad z_2(t_i) = x(t_i)\sin t_i + v_2(t_i)$$

where the sequences of $v_1(\cdot, \cdot)$ and $v_2(\cdot, \cdot)$ are independent white Gaussian sequences, each independent of $w_d(\cdot, \cdot)$ with statistics

$$E\{v_1(t_i)\} = 0, \qquad E\{v_1^2(t_i)\} = 1$$
$$E\{v_2(t_i)\} = 0, \qquad E\{v_2^2(t_i)\} = \cos^2(t_i)$$

(a)   Suppose you have an optimal estimate of the state at some time t_{i-1}, based upon the measurement history up through time t_{i-1}, and this estimate is $\hat{x}(t_{i-1}^+)$ with associated error variance $P(t_{i-1}^+)$. Write the equations for propagating the optimal estimate and error variance to the next sample time before the next measurement is taken, i.e., to obtain $\hat{x}(t_i^-)$ and $P(t_i^-)$. Explain your logic fully. If $\hat{x}(t_{i-1}^+) = 4$ and $P(t_{i-1}^+) = 1$, what are $\hat{x}(t_i^-)$ and $P(t_i^-)$?

(b)   Is there redundancy in the phrase "independent white" Gaussian sequences in the problem statement?

(c)   Let the measured values at t_i be

$$z_1(t_i, \omega_j) = z_{i1} = 3, \qquad z_2(t_i, \omega_j) = z_{i2} = 1$$

Show explicitly that the same $\hat{x}(t_i^+)$ and $P(t_i^+)$ are obtained by recursively updating the estimate with $z_1(t_i)$ and then $z_2(t_i)$ and by "batch processing" $z_1(t_i)$ and $z_2(t_i)$ simultaneously by defining a vector

$$z(t_i) = \begin{bmatrix} z_1(t_i) \\ z_2(t_i) \end{bmatrix}$$

and updating. Think before brute force evaluations—it can save significant time and algebra in obtaining the "batch process" result.

(d)   In part (c), a covariance matrix $\mathbf{R}(t_i)$ associated with the measurement noise was generated. What property of this matrix was critical to the equality of the two processing methods?

5.12   The performance function for the altitude hold mode of an airplane autopilot is given by the transfer function

$$\frac{h(s)}{h_c(s)} = \frac{0.3(s + 0.01)}{(s^2 + 0.006s + 0.003)}$$

where h represents altitude and h_c is commanded altitude. The altitude command h_c is modeled as a constant h_{c_0} plus white Gaussian noise $\delta h_c(t)$ in the command channel

$$h_c(t) = h_{c_0} + \delta h_c(t)$$

The constant h_{c_0} is modeled as a normal random variable with statistics

$$\text{Mean} = 10{,}000 \text{ ft}, \qquad \text{Variance} = 250{,}000 \text{ ft}^2$$

Noise in the command channel has the following statistics:

$$E[\delta h_c(t)] = 0, \qquad E[\delta h_c(t)\delta h_c(t + \tau)] = N_c \delta(\tau), \qquad N_c = 400 \text{ ft}^2 \text{ sec}$$

and $\delta h_c(\cdot, \cdot)$ is independent of all other processes.

Continuous measurements of altitude are available and we wish to process them to obtain the minimum variance estimate of altitude. The altitude measurements contain white noise, so the model is

$$h_m(t) = h(t) + \delta_m(t)$$

where $h_m(t)$ is measured altitude and $\delta_m(t)$ is independent white noise:

$$E[\delta_m(t)] = 0, \qquad E[\delta_m(t)\delta_m(t + \tau)] = N_m \delta(\tau). \qquad N_m = 900 \text{ ft}^2 \text{ sec}$$

(a) Determine the differential equations defining the minimum variance estimator of $h(t)$. Write these equations out in scalar form. Explain how you would determine the coefficients of these equations.

(b) Repeat, but use discrete-time measurements every second, with zero-mean white Gaussian noise of strength 900 ft^2 corrupting the measurements.

5.13 A radiometric area correlation guidance (RACG) system establishes a "position measurement" by taking a radiometric "picture" of the terrain directly below a vehicle and comparing (correlating) this with a prestored coordinatized map of the terrain. For simplicity, assume that one-dimensional position time propagation can be described by

$$x(t_{i+1}) = x(t_i) + u\,\Delta t + w_d(t_i)$$

with u being a nominal vehicle velocity, $\Delta t = (t_{i+1} - t_i) = \text{constant}$, $w_d(\cdot, \cdot)$ being zero-mean white Gaussian noise with $E\{w_d^2(t_i)\} = Q_d = \text{constant}$, and $x(t_0)$ Gaussian with mean $\hat{x}_0$ and variance P_0. Assume that the "position measurement" at time t_i is well modeled as

$$z(t_i) = x(t_i) + v(t_i)$$

with $v(\cdot, \cdot)$ zero-mean white Gaussian noise with $E\{v^2(t_i)\} = R = \text{constant}$. Assume $x(t_0)$, $w_d(\cdot, \cdot)$, and $v(\cdot, \cdot)$ are independent of each other.

Let $\Delta t = 1$ min, and assume you want to minimize the rms error in position estimate at the end of a 10-min flight. Measurements are admissible at $t_i = 0, 1, \ldots, 9$ (not at $t_i = 10$), but because prestored maps consume significant computer memory, only two maps and thus only two measurements can actually be taken. The question is, where in the flight should they be scheduled?

(a) Let $P_0 = Q_d = R = (100 \text{ ft})^2$ and solve for the optimal rms terminal position error.

(b) Assume that if the rms position error at time of measurement update should exceed 250 ft, there is an unacceptably large probability that the true vehicle position is beyond the boundaries of the prestored map, precluding a valid position measurement at all. Solve the problem again under this additional constraint.

(c) Solve part (b) for $P_0 = (300 \text{ ft})^2$, $Q_d = R = (100 \text{ ft})^2$.

(d) Solve part (b) for $Q_d = (300 \text{ ft})^2$, $P_0 = R = (100 \text{ ft})^2$.

(e) Solve part (b) for $R = (300 \text{ ft})^2$, $P_0 = Q_d = (100 \text{ ft})^2$.

5.14   Show that the inverse covariance form estimator for the case of $\mathbf{Q}_d(t_i) \equiv \mathbf{0}$ for all t_i and $\mathbf{P}_0^{-1} = \mathbf{0}$ reduces to the classical solution for the linear, unbiased, minimum variance estimate of $x(t_i)$ [assuming $\mathbf{v}(\cdot, \cdot)$ is white but not necessarily Gaussian] given by the Gauss–Markov theorem:

$$\hat{\mathbf{x}}_{G-M}(t_i) = \mathscr{I}^{-1}(t_i, t_1) \sum_{j=1}^{i} \mathbf{\Phi}^T(t_j, t_i) \mathbf{H}^T(t_j) \mathbf{R}^{-1}(t_j)$$

$$\cdot \left\{ \mathbf{z}_j + \mathbf{H}(t_j) \sum_{k=j+1}^{i} \mathbf{\Phi}(t_j, t_k) \mathbf{B}(t_{k-1}) \mathbf{u}(t_{k-1}) \right\}$$

5.15   Show that the optimal *prediction* of $\mathbf{x}(t_j)$ based on measurements through time $t_i < t_j$, $E\{\mathbf{x}(t_j) | \mathbf{Z}(t_i) = \mathbf{Z}_i\}$, can be evaluated by means of

$$\hat{\mathbf{x}}(t_j | t_i) = \mathbf{\Phi}(t_j, t_i)\hat{\mathbf{x}}(t_i^+) + \int_{t_i}^{t_j} \mathbf{\Phi}(t_j, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$$

$$\mathbf{P}(t_j | t_i) = \mathbf{\Phi}(t_j, t_i)\mathbf{P}(t_i^+)\mathbf{\Phi}^T(t_j, t_i) + \int_{t_i}^{t_j} \mathbf{\Phi}(t_j, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\mathbf{\Phi}^T(t_j, \tau)\,d\tau$$

Show that this can be generated recursively by iterating only the time propagation relations of a Kalman filter, without measurement updates, $(j - i)$ times from the initial conditions $\hat{\mathbf{x}}(t_i^+)$ and $\mathbf{P}(t_i^+)$.

One practical use of this idea is to partition a sample period Δt into N subintervals to maintain accuracy in numerical integration. Then the filter iteration period is $(\Delta t/N)$, and a measurement update is computed only every N propagations, generating optimal predictions at each intermediate point. Similarly, if one measurement becomes available every $N \Delta t$ sec and another every $M \Delta t$ sec, with M and N unequal integers, the filter iteration period could be set at a constant Δt sec, providing optimal state predictions at points where neither measurement becomes available.

5.16   Show that the Joseph form covariance measurement update for scalar measurements can be expressed equivalently in the computationally efficient manner:

$$\mathbf{a} = \mathbf{P}(t_i^-)\mathbf{H}^T(t_i), \qquad\qquad \mathbf{P}_1 = \mathbf{P}(t_i^-) - \mathbf{K}(t_i)\mathbf{a}^T$$

$$\mathbf{b} = \mathbf{P}_1\mathbf{H}^T(t_i) - \mathbf{K}(t_i)R(t_i), \qquad \mathbf{P}(t_i^+) = \mathbf{P}_1 - \mathbf{b}\mathbf{K}^T(t_i)$$

5.17   Let a signal of interest be the output of a first order lag driven by white Gaussian noise $w_1(\cdot, \cdot)$, and let that signal be corrupted by exponentially time-correlated noise $n(\cdot, \cdot)$, modeled as the output of a first order shaping filter driven by white Gaussian noise $w_2(\cdot, \cdot)$, as depicted in Fig. 5.P2. The system (plant) is described by the transfer function $F_p(s) = 1/(s + \omega_p)$ and the noise shaping filter is described by $F_n(s) = 1/(s + \omega_n)$. Relevant statistics are

$$E\{w_1(t)\} = 0, \qquad E\{w_1(t)w_1(t + \tau)\} = 2\,\delta(\tau)$$

$$E\{w_2(t)\} = 0, \qquad E\{w_2(t)w_2(t + \tau)\} = 1\,\delta(\tau)$$

Assume $w_1(\cdot, \cdot)$ and $w_2(\cdot, \cdot)$ are independent and that the appropriate initial conditions are

$$E\{s(t_0)\} = 0, \qquad E\{n(t_0)\} = 0 \qquad E\{s^2(t_0)\} = 1 \qquad E\{n^2(t_0)\} = \tfrac{1}{2}$$

Determine a recursion equation for the variance of the error in the estimate of the signal using discrete-time measurements $z(t_i)$. By performing a coordinate transformation, you should be able
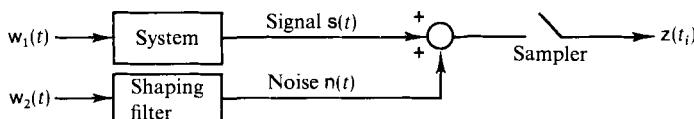
FIG. 5.P2   System schematic for Problem 5.17.

to express the recursion for the appropriate element of $\mathbf{P}^*(t_i{}^+)$ solely in terms of the value of that same element at time t_{i-1}.

5.18   Consider the first order system modeled by

$$\dot{x}(t) = w(t)$$

where $w(\cdot,\cdot)$ is white Gaussian noise with statistics

$$E[w(t)] = 0, \qquad E[w(t)w(t+\tau)] = 4\,\delta(\tau)$$

Assume that at time $t = 0$, the initial state $x(0)$ is modeled as a Gaussian random variable with statistics

$$E[x(0)] = 10, \qquad E[\{x(0) - 10\}^2] = 25$$

The observed signal is

$$z(t) = x(t) + v(t)$$

where $v(\cdot,\cdot)$ is white Gaussian noise independent of $w(\cdot,\cdot)$ with

$$E[v(t)] = 0, \qquad E[v(t)v(t+\tau)] = 16\,\delta(\tau)$$

First determine the exact equations of the Kalman filter to estimate $x(t)$ for all t. Then investigate the steady state behavior of the filter as $t \to \infty$. Show that this steady state behavior is in fact time invariant; determine its transfer function.

5.19   Given the linear system model depicted in Fig. 5.P3, where $w_1(\cdot,\cdot)$, $w_2(\cdot,\cdot)$, $v(\cdot,\cdot)$, $x_1(0)$, and $x_2(0)$ are mutually independent, zero mean, and Gaussian, with

$$E\{w_1(t)w_1(t+\tau)\} = \delta(\tau), \qquad E\{w_2(t)w_2(t+\tau) = \delta(\tau)$$
$$E\{x_1(0)^2\} = 1, \qquad E\{x_2(0)^2\} = 2, \qquad E\{v(t)v(t+\tau) = 2\delta(\tau)$$

Determine the optimal estimator for $x_3(t)$ using continuous measurements $z(t)$. (Question to answer first: How many state variables are required to describe the system?)
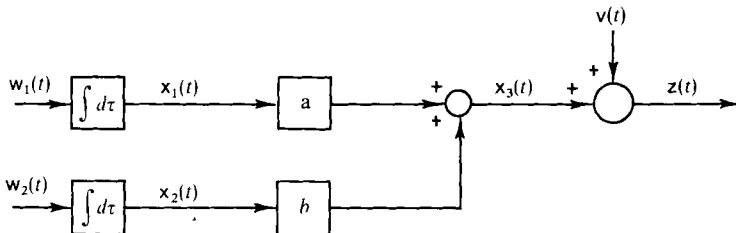
FIG. 5.P3   System model for Problem 5.19.

5.20   Recall the satellite orbit problem discussed in Examples 2.7 and 2.8. The perturbation equations for describing small deviations from an assumed circular orbit were given in Example 2.8. Note that there is no process noise included in these state equations. It is desired to measure these small orbital deviations from observations on the ground. Two proposals are presented:

(1)   In an effort to keep the measurement stations rather simple and inexpensive, only angle (x_3) measurements will be made. However, the designer realizes the very likely possibility of measurement errors and includes an optimal filter in his proposal for estimating the states. The measurement may be represented as

$$z_1(t) = x_3(t) + v_1(t)$$

where

$$E[v_1(t)] = 0, \qquad E[v_1(t)v_1(t + \tau)] = R_1\,\delta(\tau)$$

(2) The second design proposes to use measurements of range (x_1) only. In this case,

$$z_2(t) = x_1(t) + v_2(t)$$

where

$$E[v_2(t)] = 0 \qquad E[v_2(t)v_2(t + \tau)] = R_2\,\delta(\tau)$$

It is your task to determine which of these proposals is superior. Are both system models observable? What does that indicate? Is there any benefit to incorporating both z_1 and z_2?

5.21 A linear system has the input/output transfer function

$$\frac{x(s)}{w(s)} = \frac{s + \alpha}{s + \beta}$$

The input $w(t)$ is known to be identically zero. The initial condition on the system, $x(0)$, is modeled as a Gaussian random variable with

$$E\{x(0)\} = 0, \qquad E\{x(0)^2\} = 1$$

Continuous measurements of the form

$$z(t) = x(t) + v(t)$$

are available with $v(\cdot,\cdot)$ a white Gaussian noise with

$$E\{v(t)\} = 0, \qquad E\{v(t)v(t + \tau)\} = \delta(\tau)$$

Determine the optimal estimator of $x(t)$ and the associated error variance explicitly for all $t \geq 0$.

To obtain an explicit evaluation of the variance as a function of time, the following fact can be useful. If the positive definite matrix **M** satisfies the differential equation

$$\dot{\mathbf{M}} = \mathbf{AM} + \mathbf{MA}^{\mathrm{T}} - \mathbf{MBM}$$

then $\mathbf{M}^{-1}$ satisfies the linear equation

$$\dot{\mathbf{M}}^{-1} = -\mathbf{A}^{\mathrm{T}}\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{A} + \mathbf{B}$$

5.22 Consider the unstable first order system

$$\dot{x}(t) = x(t) + w(t)$$

with measurements

$$z(t) = x(t) + v(t)$$

where

$$E\{w(t)\} = 0, \qquad E\{w(t)w(t + \tau)\} = Q\,\delta(\tau)$$
$$E\{v(t)\} = 0, \qquad E\{v(t)v(t + \tau)\} = R\,\delta(\tau)$$

Let $Q = 1$.

(a) Solve for the error variance, $P(t)$, as a function of time, assuming $P(t = 0) = P_0$. Using your expression for $P(t)$, evaluate $\lim_{t \to \infty} P(t)$. For values of $R = 1, 2,$ and 4, sketch the curve $P(t)$ as a function of time. (Use $P_0 = 1$.)

(b) Consider the homogeneous part of the filter which is given by the differential equation

$$\dot{y} = [1 - P/R]\,y$$

Using the Lyapunov function $V(\cdot,\cdot)$ defined by $V(y,t) = y^{\mathsf{T}}P^{-1}(t)y$, the system can be shown to be asymptotically stable. Why is this significant?

5.23 A continuous measurement process, $z(\cdot,\cdot)$ is given as

$$z(t) = at + n(t)$$

where a is modeled as a Gaussian random variable with

$$E[a] = 0, \qquad E[a^2] = 1$$

and $n(\cdot,\cdot)$ is a white Gaussian noise process with

$$E[n(t)] = 0, \qquad E[n(t)n(t + \tau)] = 2\,\delta(\tau)$$

Obtain the optimal filter for estimating a. Is the filter a stable system?

5.24 In Example 5.16 of Section 5.12, it was stated that for small signal-to-noise ratio, $c \cong a$ and the low frequency gain is about $\frac{1}{2}[A/(\Psi_0 a^2)]$. Show this.

5.25 Show that Eq. (5-171) is valid.

5.26 Consider a system described by $F_p(s) = 1/s$ driven by white Gaussian noise $w_1(\cdot,\cdot)$ whose output is corrupted by exponentially time-correlated noise described as the output of the noise shaping filter $F_n(s) = 1/(s + a)$ driven by white Gaussian noise $w_2(\cdot,\cdot)$. See Fig. 5.P4. Noise statistics of the uncorrelated w_1 and w_2:

$$E[w_1(t)] = 0, \qquad E[w_1(t)w_1(t + \tau)] = Q_1\,\delta(\tau)$$
$$E[w_2(t)] = 0, \qquad E[w_2(t)w_2(t + \tau)] = Q_2\,\delta(\tau)$$

initial condition statistics:

$$E\{x_1(t_0)\} = 0, \qquad E\{x_2(t_0)\} = 0$$
$$E[x_1(t_0)^2] = \sigma_1{}^2, \quad E[x_2(t_0)^2] = \sigma_2{}^2, \quad E[x_1(t_0)x_2(t_0)] = 0$$
$$E[z(t_0)^2] = \sigma_z{}^2, \qquad E[z(t_0)x_1(t_0)] = \sigma_1{}^2$$

Determine the optimal continuous-time estimate for the state variable x_1. Observe the behavior as $t \to \infty$. Show that it can be described by a time-invariant system with transfer function

$$\frac{\hat{x}_1(s)}{z(s)} = \frac{c_1(s + a)}{s + ac_1} \qquad c_1 = \sqrt{\frac{Q_1}{Q_1 + Q_2}}$$
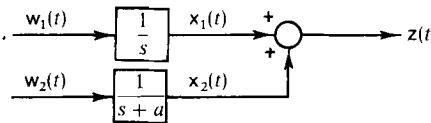
FIG. 5.P4 System configuration for Problem 5.26.

5.27 Show that the result of the previous problem is identical to the result of a Wiener filter design for the steady state problem.

5.28 Design an optimum linear filter (Wiener filter) to separate noise $n(\cdot,\cdot)$ from a signal $s(\cdot,\cdot)$ when these processes are uncorrelated with each other and

$$\Psi_{ss}(\omega) = 1/(\omega^2 + 1), \qquad \Psi_{nn}(\omega) = 2\omega^2/(\omega^4 + 1)$$

5.29 (a) An engineer presents you with a single input/single output black box that he says contains a steady state Kalman filter. To test the performance of the filter you subject it to a time-correlated noise with autocorrelation function

$$\Psi_{nn}(\tau) = E\{n(t)n(t + \tau)\} = Ne^{-5|\tau|}$$

and obtain an output whose power spectral density is

$$\Psi_{yy}(\omega) = \frac{\frac{5}{2}N(\omega^2 + 16)}{(\omega^2 + 4)(\omega^2 + 25)}$$

What do you think of the engineer's competence?

(b) Somewhat perplexed, you go back and ask him how he designed the filter. He tells you that he was faced with design of a filter to separate a signal from a signal-plus-noise input, where the signal power spectral density $\Psi_{ss}(\omega)$ and noise power spectral density $\Psi_{nn}(\omega)$ could be approximated as

$$\Psi_{ss}(\omega) = \frac{5/12}{\omega^2 + 4}, \qquad \Psi_{nn}(\omega) = \frac{7/12}{\omega^2 + 16}$$

and the signal and noise are not correlated with each other. Calculate the Wiener optimum filter, and compare your answer to his, and offer him any appropriate constructive criticism.

5.30 Show that the discrete-time Kalman filter algorithm of Table 5.2 can also be expressed in the following form ("innovations form") [32, 33]:

$$\hat{\mathbf{x}}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\hat{\mathbf{x}}(t_i^-) + \mathbf{B}_d(t_i)\mathbf{u}(t_i) + \mathcal{K}(t_i)\mathbf{v}(t_i)$$
$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) + \mathbf{\Sigma}(t_i)\mathbf{v}(t_i)$$

where

$$\mathbf{\Sigma}(t_i) = [\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^\mathsf{T}(t_i) + \mathbf{R}(t_i)]^{1/2}$$
$$\mathbf{v}(t_i) = \mathbf{\Sigma}^{-1}(t_i)[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$$
$$\mathcal{K}(t_i) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i^-)\mathbf{H}^\mathsf{T}(t_i)\mathbf{\Sigma}^{-\mathsf{T}}(t_i)$$
$$\mathbf{P}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i^-)\mathbf{\Phi}^\mathsf{T}(t_{i+1}, t_i) + \mathbf{G}_d(t_i)\mathbf{Q}_d(t_i)\mathbf{G}_d^\mathsf{T}(t_i) - \mathcal{K}(t_i)\mathcal{K}^\mathsf{T}(t_i)$$

Note that the square root matrix used in the preceding is defined such that $\mathbf{A}^{1/2}(\mathbf{A}^{1/2})^\mathsf{T} = \mathbf{A}$ for a given matrix $\mathbf{A}$. Also show that $\mathbf{v}(\cdot, \cdot)$ is a zero-mean white Gaussian sequence with $E\{\mathbf{v}(t_i)\mathbf{v}^\mathsf{T}(t_j)\} = \mathbf{I}\delta_{ij}$.

Design and performance analysis
of Kalman filters

6.1 INTRODUCTION

This chapter seeks to exploit the Kalman filter algorithm to its fullest potential. To do so will encompass a complete depiction of a systematic design procedure, practical aspects of implementation, and development of software tools to provide performance analysis capability for any Kalman filter configuration operating in the real world environment. Extensive examples are the vehicle for developing and emphasizing the essential points of designing efficient, practical filters, and these examples are drawn from the application area that has probably exploited Kalman filtering the most: optimally aided inertial navigation systems. This is by no means the only important applications area, but concentrating attention on one particular problem area will allow a more extensive portrayal of filter design than would a superficial look at many different contexts. For additional applications, see Leondes [28] and Schmidt [47].

6.2 THE REQUISITE OF ENGINEERING JUDGMENT

Chapter 5 may have left the impression that an optimal filter can be generated automatically once a body of applied mathematics has been mastered. Despite the mathematical formalism of the Kalman filter approach, a substantial amount of engineering insight and experience is required to develop an effective operational filter algorithm.

As mentioned in Chapter 5, a mathematical model of both the system structure (state dynamics and output relations) and uncertainties is inherently embodied in the Kalman filter structure. Attaining an *adequate mathematical model upon which to base the filter* is the crux of the design problem. Even once

289

a particular model is chosen as appropriate for a given application, a considerable effort remains: obtaining appropriate *numerical evaluation* of coefficients (especially covariance matrix elements) within the model. This process is called "tuning" a given Kalman filter, and it involves an iterative search for the coefficient values that yield the best estimation performance possible from that particular filter structure.

Moreover, the design must meet the constraints of online computer time, memory, and wordlength required. These considerations dictate a philosophy of using as simple a filter as possible that yields adequate results (i.e., meeting performance specifications). Consequently, the designer must be able to exploit basic modeling alternatives to achieve a simple but adequate filter, adding or deleting model complexity as the performance needs and practical constraints require.

Evaluation of *true* performance capabilities of simplified, reduced order filters is of critical importance in the design procedure. Although a Kalman filter computes an error covariance internally, this is a valid depiction of the true errors committed by the filter only to the extent that the filter's own system model adequately portrays true system behavior. It is very possible for the filter not to perform as well as it "thinks" it does. If the computed error covariance is inappropriately small, so is the computed gain: the filter weights its internal system model too heavily and discounts the data from the "real world" too much, leading to filter estimates not corresponding to true system performance, with a simultaneous indication by the filter through its computed covariance that all is well, a condition called filter divergence.

Moreover, numerical precision and stability problems can corrupt performance substantially, especially when the filter is implemented on a short wordlength computer. This motivates consideration of alternative algorithm formulations, such as square root filters (to be discussed in Chapter 7).

Evaluation of true performance capabilities involves both covariance analyses and Monte Carlo simulations, which will be discussed in Section 6.8. To achieve a valid portrayal, the designer must fully understand the assumptions that underly a statistical analysis of performance. He must investigate the effects of nonwhite noises, non-Gaussian noises, neglected nonlinearities, and approximations used in achieving a model form compatible with the Kalman filter assumptions. Moreover, testing and simulation experience with the actual digital implementation is crucial before the software is finalized.

In certain applications, the basic Kalman filter structure itself has to be modified. For example, sensor data (rather noisy) is often available more frequently than the established sample rate of the filter algorithm (especially since most sensors are analog devices). In this case, one might consider prefiltering of the data to smooth out the noise, rather than just sample it periodically and neglect the information available between sample instants. One can in fact generate an equivalent single data value that incorporates this information,

but it is not the simple replacement of raw data by an averaged signal that first occurs to an engineer (this will be pursued later in Section 6.5). Other examples would be artificially limiting the filter's "memory" of past data, setting lower bounds on acceptable variance evaluations, discarding measurement samples that fail reasonableness tests, and adaptively setting filter gains.

The design of an effective operational Kalman filter entails an iterative process of proposing alternative designs through physical insights, tuning each, and trading off performance capabilities and computer loading. A systematic procedure for accomplishing such a design, as developed in Section 6.9, will accentuate the need to bring engineering judgment to bear on the overall filter development.

6.3   APPLICATION OF KALMAN FILTERING TO INERTIAL NAVIGATION SYSTEMS

To consider basic aspects of filter implementation in the context of realistic applications, the next sections of this chapter will concentrate upon applying a Kalman filter to inertial navigation systems. Fundamental concepts will be developed in this section, followed by three specific examples in succeeding sections.

A conventional gimbaled inertial measurement unit [2, 4, 25, 40, 41] consists of a platform suspended by a gimbal structure that allows three degrees of rotational freedom, as depicted in Fig. 6.1. From the geometry of the configuration, it is possible to attach the outermost gimbal to the body of some vehicle, and to allow that vehicle to undergo any change in angular orientation while maintaining the platform fixed with respect to some desired coordinate frame. Gyros on the platform sense the angular rate of the platform with respect to inertial space, and their outputs are sent through electronics to the torquer motors on the gimbal structure, commanding them to maintain a desired platform orientation regardless of the orientation of the outermost gimbal. Feedback control loops that keep the gyro outputs nulled will maintain the platform fixed with respect to inertial space; additional computed inputs can be added to these loops to maintain some other orientation, such as north–east–down corresponding to the current location of the vehicle. These feedback loops are such that, in practice, the platform orientation is kept essentially stable regardless of the most violent vehicle maneuvering.

Thus the platform remains aligned with respect to a known reference coordinate system. Accelerometers on the platform then provide vehicle acceleration with respect to that known set of reference coordinates. Actually, accelerometers measure specific force, so local gravity must be computed and subtracted appropriately from these sensor outputs to obtain a measurement of actual vehicle acceleration. The resulting signals can be integrated (or pulses counted

Gimbal two

Controlled member
platform
(gimbal three)

M_3

ϕ

Gimbal one

M_2

Gyro

Accelerometers

Torquer
motor

M_1

θ

ψ

Base aircraft
(gimbal zero)

M_1 M_2 M_3 x y
ϕ θ ψ
                                                                        z

Gimbal zero   Gimbal one   Gimbal two   Gimbal three
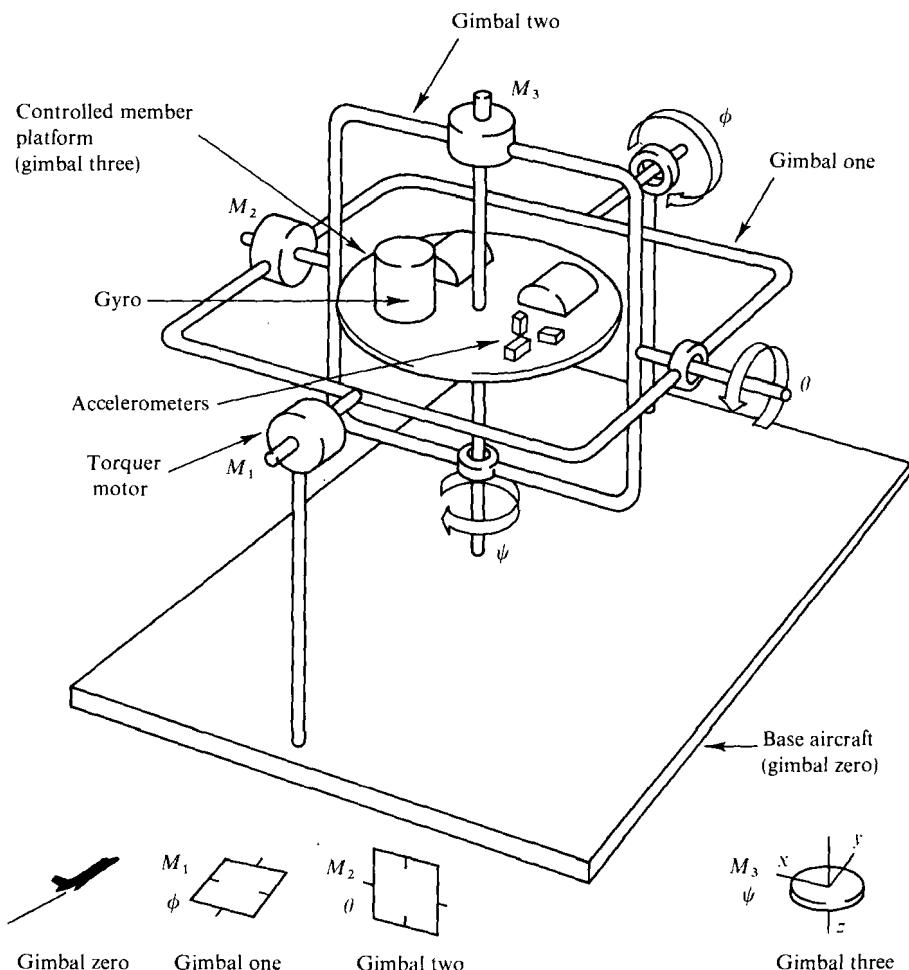
FIG. 6.1   Typical gimbaled inertial measurement unit. Modified from *Gyroscopic Theory, Design and Instrumentation* by Wrigley, Hollister, and Denhard. © 1969. Used with the permission of The M.I.T. Press.

if the signal format is a pulse rate proportional to acceleration) to yield vehicle velocity and position.

Moreover, the gimbal angles, the angles formed between the various members of the gimbal support structure, provide direct readout of the Euler angles to describe the vehicle's angular orientation. Thus, the inertial navigation system (INS) provides attitude information as well as translational information.

The question naturally arises, why does this instrument require optimal aiding by other navigation sensors? Due to the tight control loops just described, an INS provides very good high frequency information. However, because of

gyro characteristics, the system drifts at a slow rate: the long term, or low frequency, content of the data is poor. All inertial systems have position errors that grow slowly with time, and these errors are unbounded. A typical INS might be classed as a "one nautical mile per hour system," meaning that after one hour of operation, the position error standard deviation has grown to one nautical mile.

Some external source of data, such as radio navigation aid position information or Doppler velocity, would naturally be considered as a means of bounding or damping these errors. As opposed to an INS, most other navigation aids provide data which is good on the average (i.e., low frequency), but subject to considerable high frequency noise, due to instrument noise, atmospheric effects, antenna oscillation, unlevel ground effects, and so forth.

Given an INS and other sources of data, one would want to combine the available information in an optimal manner if possible, efficiently providing an estimate of navigation parameters that is best with respect to some criterion. Some earlier navigation systems reset the INS to agree with the other data sources, essentially declaring these other sources perfect and discarding the information previously held in the INS. The Kalman filter approach is instead to use the statistical characteristics of the errors in both the external information and the inertial components to determine this "optimal" combination of information. Actually, the filter *statistically* minimizes the errors in the estimates of the navigation parameters: on an ensemble average basis, no other means of combining the data will outperform it, *assuming* the internal model in the filter is adequate. Once the problem and system model are completely specified, the Kalman filter algorithm systematically provides this optimal estimate.

It will be seen that, although the filter is designed in the time domain, it does in fact weight each information source most heavily in the frequency regime where it provides good information and suppress each in the frequency region where it is most prone to errors. In other words, the filter will use the good low frequency data from the external sources to damp out the slowly growing errors inherent in the INS. Because of these differing sensor characteristics and the existence of well-developed adequate system models, there is substantial benefit to be gained by applying Kalman filtering to aiding INS systems.

Typical *external information sources* would include

Position data:

(1) radar—onboard and/or ground based;
(2) radio navigation aids: TACAN, LORAN, OMEGA, VOR/DME;
(3) Global Positioning System (GPS) navigation satellites;
(4) position fixes: star sightings, landmarks;
(5) radiometric area correlation (comparison of a radiometric "picture" of terrain to a stored map);
(6) laser ranging.

Velocity data:

(1) Doppler radar,
(2) indicated airspeed from the air data system.

Altitude data:

(1) barometric altimeter,
(2) radar altimeter,
(3) laser altimeter.

The filter for combining these data sources with an INS is typically a digital computer algorithm that uses sampled data (with sample period on the order of 5–60 sec) to maintain estimates of approximately 10–20 state variables. For instance, the navigation system filter designed for the B-1 bomber (which will serve as a basis for future designs) consists of a 13-state horizontal plane subsection and an independent 4-state vertical channel component, operating with a 6-sec sample period.
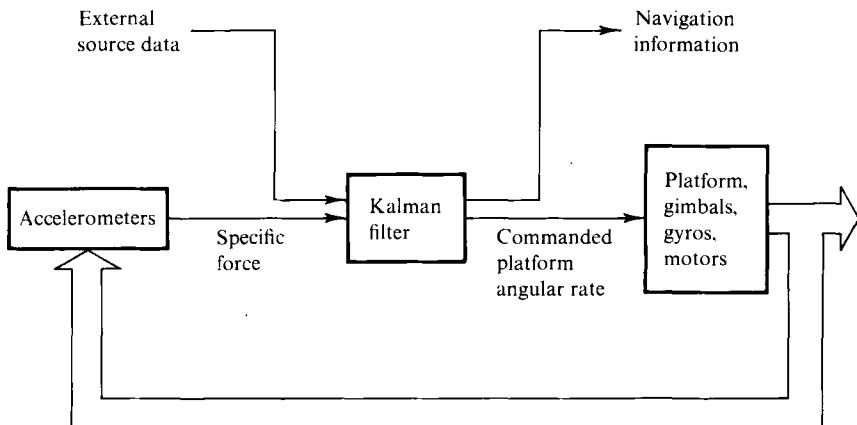
There are two very important aspects of implementation of a Kalman filter in conjunction with inertial systems (and other applications): *total state space versus error state space formulation* (also denoted as *direct* versus *indirect* filtering in navigation literature), and *feedforward versus feedback mechanizations* [3]. These aspects will now be discussed.

As the name indicates, in the *total state space (direct)* formulation, total states such as vehicle position and velocity are among the state variables in the filter, and the measurements are INS accelerometer outputs and external source signals. In the *error state space (indirect)* formulation, the *errors* in the INS-indicated position and velocity are among the estimated variables, and each measurement presented to the filter is the *difference* between INS and external source data.

Consider first the *total state space filter*, as depicted in Fig. 6.2. In this direct configuration, the Kalman filter is *in* the INS loop. The INS accelerometer signals and external source data both feed into the filter, which provides not only the desired navigation information, but also the appropriate commands to the gimbal torquer motors to maintain the platform aligned with the chosen coordinate frame. The angular orientation of the platform then dictates what accelerations will be sensed by each of the accelerometers attached to it.

The benefits of such a configuration are that the available information is weighted optimally rather than operated upon by fixed gains and integrators. Optimal time-varying gains can provide substantial improvement in performance over the classical approaches of resets or fixed gain updating, such as reducing INS gyrocompass (initial alignment) time from about 30 to 7 min while meeting the same precision specification.

However, there is a very serious drawback to this implementation. Being in the INS control loop and using the total state space representation, the filter

Physical connection by platform angular orientation

FIG. 6.2   Total state space (direct) Kalman filter.

would have to maintain explicit, accurate awareness of vehicle angular motion as well as attempt to suppress noisy and erroneous data. Sampled data usage requires a sampling rate of at least twice the frequency of the highest frequency signal of interest for adequate reconstruction of the continuous-time system behavior (by the Shannon sampling theorem; engineering practice tends toward five to ten times the highest frequency signal). Admitting aircraft roll rate capability on the order of 400 deg/sec, the filter would need a very fast sample rate and would have to perform all computations within this short sample period. Moreover, in most cases, the Kalman filter is allocated only a small portion of the capabilities of a central processor, and it is run "in the background" at a lower priority than more critical algorithms, such as digital stability and control programs. It is impossible to implement a high order Kalman filter practically on state-of-the-art computers and meet this sample rate requirement.

Not only are the dynamics involved in the total state space description composed of high frequency content, but they are well described only by a nonlinear model. The development of a Kalman filter is predicated upon an adequate linear system model, and such a total state space model does not exist.

Another drawback to this design is that, if the filter should happen to fail (as by a temporary computer failure), the entire navigation system fails: the inertial system *cannot* operate without the filter. From a reliability standpoint, it would be desirable to provide an emergency degraded performance mode in case of such failure.

As a result of these considerations, the direct mechanization is restricted to alignment, calibrations, bias determination in laboratory testing, certain submarine applications (involving slower dynamics), and the like. Section 6.6 will pursue this subject further.

The *error state space (indirect) Kalman filter* estimates the errors in the navigation and attitude information using the difference between INS and external source data. The INS itself follows the high frequency motions of the vehicle very accurately, and there is no need to model these dynamics explicitly in the filter. Instead, the dynamics upon which the filter is based is the set of inertial system error propagation equations, which are relatively well developed, well behaved, low frequency (Schuler 84-min mode dominant), and very adequately represented as linear. Because the filter is out of the INS loop and is based on low frequency linear dynamics, its sample rate can be much lower than that of a direct filter. In fact, an effective indirect filter can be developed with a sample period on the order of half a minute, thereby achieving practicality with respect to the amount of computer time required. For these reasons, the error state space formulation is used in essentially all (except submarine) terrestrial aided inertial navigation systems.

Besides state space differences among Kalman filters, there are two distinct types of implementations, feedforward and feedback. The *indirect feedforward* version is depicted in Fig. 6.3. From this diagram, it can be seen that the filter compares the two sets of data and uses the result to estimate the errors in the inertial system. By subtracting these estimated errors from the inertial data, the onboard computer maintains the optimal estimates of position, velocity, and attitude. The inertial system operates as though there were no aiding: it is "unaware" of the existence of the filter or the external data, so if either should fail, the unaltered INS information would still be available. Herein lies the disadvantage of the feedforward approach though. Acceptable Kalman filter performance depends upon the adequacy of a *linear* dynamics model, so it is necessary for the errors in the inertial system to remain of small magnitude. However, the INS is free to drift with unbounded errors, thereby invalidating this basic assumption.
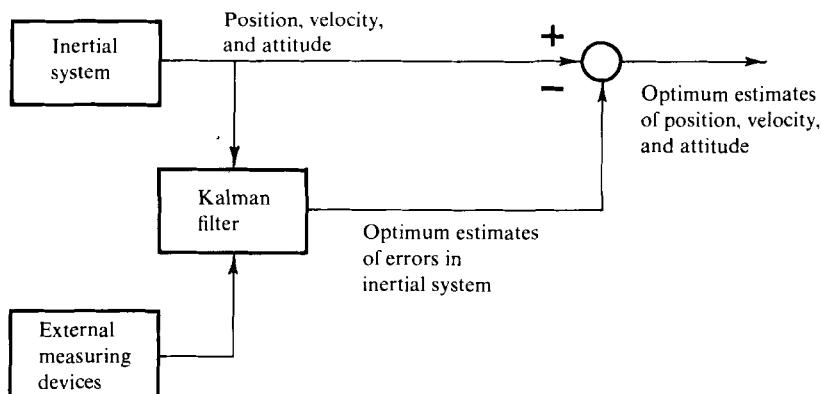
FIG. 6.3   Indirect feedforward Kalman filter.
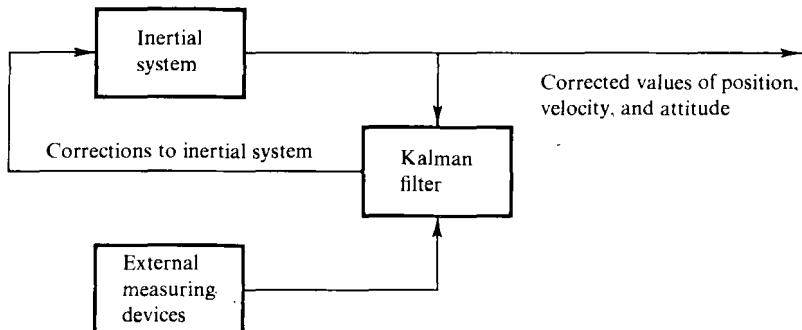
FIG. 6.4   Indirect feedback Kalman filter.

Thus, an *indirect feedback* configuration as in Fig. 6.4 is motivated. The Kalman filter again generates estimates of the errors in the inertial system, but these are fed back into the INS to correct it. In this way, the inertial errors are not allowed to grow unchecked, and the adequacy of a linear model is enhanced. There is a further advantage to this configuration. Since the INS is corrected after each measurement sample, many of the predicted error states at the next sample time will be zero, and thus these components of $\hat{x}(t_{i+1}^-)$ need not be computed explicitly. With regard to filter or external aid failure, because of the slow sample rate and slow INS error dynamics, such failures could be detected and the corrections to the INS could be removed before much (any) performance deterioration were caused.

The general comments of this section will be developed further in the following sections, which consider two error state space filters and a total state space design.

6.4   INS AIDED BY POSITION DATA: A SIMPLE EXAMPLE

Let us consider the combining of inertial system data with position data provided by radar or a radio navigation aid [21–23, 25, 30, 52]. Conceptually, we will want to weight the INS information heavily in the high frequency range (where it provides good data), and emphasize the position data in the low frequency range. This will be a very simple problem formulation confined to a single direction. More complicated, three-dimensional system models are more realistic, but the two-state model to be studied will allow simple algebra and greater transparency of effects of system model characteristics upon filter performance. The insights gained will be directly extendable to more complex problems. Furthermore, this problem will be formulated with continuous-time measurements to allow direct frequency domain interpretation of the filter's
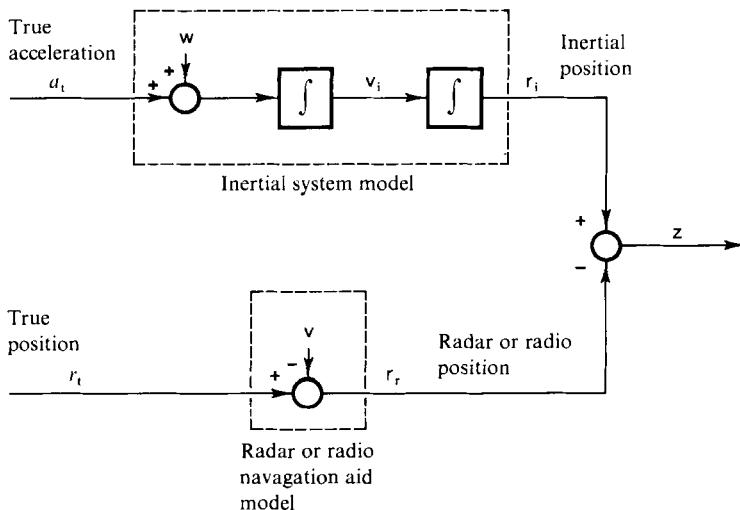
FIG. 6.5   Simple system model for INS aided by position data.

operation. The more realistic case of sampled-data measurements will be considered subsequently.

For this example, model the inertial navigation system simply as a double integrator of noise-corrupted acceleration information, as depicted in Fig. 6.5. The noise $w(\cdot,\cdot)$ is a white Gaussian noise of mean zero and variance kernel

$$E\{w(t)w(t + \tau)\} = Q\,\delta(\tau) \qquad (6\text{-}1)$$

entering at the acceleration level, and it is meant to model the errors corrupting the INS accelerometer outputs (accelerometer biases and noise, platform misalignment, etc.). The noise-corrupted acceleration is integrated once to yield INS-indicated velocity $v_i(\cdot,\cdot)$, and a second time to obtain inertially-indicated position $r_i(\cdot,\cdot)$.

Similarly, consider a simple model for the radar or radio navigation aid (a model which is actually incorporated into many operational filters) as the true position r_t, corrupted by noise $v(\cdot,\cdot)$. The $v(\cdot,\cdot)$ process is a zero-mean white Gaussian noise of strength R_c:

$$E\{v(t)v(t + \tau)\} = R_c\,\delta(\tau) \qquad (6\text{-}2)$$

meant to model the wideband noise corrupting the data provided by a radar or radio navigation aid. The negative sign in the model in Fig. 6.5 is just for convenience, as will be seen in a moment. This noise is denoted as $v(\cdot,\cdot)$ to correspond to the notation adopted earlier, and should not be confused with the subscripted velocity variables (subscripts used in this development are t = true, i = inertial system, and r = radar or radio navigation aid).

The two error state variables for this example are

$$\delta r(t) = r_i(t) - r_t(t) \tag{6-3a}$$

$$\delta v(t) = v_i(t) - v_t(t) \tag{6-3b}$$

i.e., the errors in INS-indicated position and velocity. The measurement to be presented to the filter is the difference between the inertially indicated position and that measured by the radar or radio navigation aid; from the figure,

$$z(t) = r_i(t) - r_r(t) \tag{6-4}$$

$$= [r_t(t) + \delta r(t)] - [r_t(t) - v(t)]$$

$$= \delta r(t) + v(t) \tag{6-5}$$

This then is a "measurement" of the *error* $\delta r(t)$, corrupted by noise $v(t)$; the original negative sign on $v(\cdot, \cdot)$ yields the positive sign here.

To establish the state dynamics model for the error states, first consider the total states $r_i(\cdot, \cdot)$ and $v_i(\cdot, \cdot)$. From Fig. 6.5 we can write (in white noise notation)

$$\begin{bmatrix} \dot{r}_i(t) \\ \dot{v}_i(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_i(t) \\ v_i(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [a_t(t) + w(t)] \tag{6-6}$$

However, the true position, velocity, and acceleration are related by

$$\begin{bmatrix} \dot{r}_t(t) \\ \dot{v}_t(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_t(t) \\ v_t(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} a_t(t) \tag{6-7}$$

Subtracting (6-7) from (6-6), and using the error state definitions of (6-3), yields the desired relations as

$$\begin{bmatrix} \dot{\delta r}(t) \\ \dot{\delta v}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \delta r(t) \\ \delta v(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w(t) \tag{6-8a}$$

$$(\dot{x}(t) \quad = \quad F \quad x(t) \quad + \quad G \; w(t)) \tag{6-8b}$$

In problems for which the total state model is nonlinear rather than as in (6-6), the error state equation would be derived through perturbation techniques. In terms of this state vector notation, (6-5) becomes

$$z(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \delta r(t) \\ \delta v(t) \end{bmatrix} + v(t) \tag{6-9a}$$

$$(z(t) = \quad H \quad x(t) \quad + v(t)) \tag{6-9b}$$

Initial conditions are established by modeling $x(t_0)$ as a Gaussian random variable, with mean zero (appropriate for error states) and covariance P_0:

$$P_0 = \begin{bmatrix} \sigma_r^2 & r_{rv}\sigma_r\sigma_v \\ r_{rv}\sigma_r\sigma_v & \sigma_v^2 \end{bmatrix} \tag{6-10}$$

The diagonal terms σ_r^2 and σ_v^2 are the variances of uncertainty in knowledge of initial position and velocity, respectively, and the off-diagonal term is the initial cross correlation between position and velocity, with r_{rv} the associated correlation coefficient.

The objective is to use the measured difference

$$[r_i(t, \omega_j) - r_r(t, \omega_j)] = [r_i(t) - r_r(t)] \tag{6-11}$$

for all time t of interest, combined with knowledge of the system model and statistical description of initial conditions and the noises and/or uncertainties in both data sources, to generate the optimal estimate of the state realization

$$\mathbf{x}(t, \omega_j) = \begin{bmatrix} \delta r(t, \omega_j) \\ \delta v(t, \omega_j) \end{bmatrix} = \begin{bmatrix} \delta r(t) \\ \delta v(t) \end{bmatrix} \tag{6-12}$$

The continuous-time Kalman filter for this problem is given by (5-144) to (5-146) as

$$\begin{bmatrix} \dot{\hat{\delta r}}(t) \\ \dot{\hat{\delta v}}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\delta r}(t) \\ \hat{\delta v}(t) \end{bmatrix} + \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{12}(t) & P_{22}(t) \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{1}{R_c} [z(t) - \hat{\delta r}(t)]$$

$$= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\delta r}(t) \\ \hat{\delta v}(t) \end{bmatrix} + \begin{bmatrix} P_{11}(t)/R_c \\ P_{12}(t)/R_c \end{bmatrix} [z(t) - \hat{\delta r}(t)] \tag{6-13}$$

where the covariance $\mathbf{P}(\cdot)$ satisfies

$$\begin{bmatrix} \dot{P}_{11}(t) & \dot{P}_{12}(t) \\ \dot{P}_{12}(t) & \dot{P}_{22}(t) \end{bmatrix} = \begin{bmatrix} P_{12}(t) & P_{22}(t) \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} P_{12}(t) & 0 \\ P_{22}(t) & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}$$

$$- \begin{bmatrix} P_{11}^2(t)/R_c & P_{11}(t)P_{12}(t)/R_c \\ P_{11}(t)P_{12}(t)/R_c & P_{12}^2(t)/R_c \end{bmatrix} \tag{6-14}$$

starting from the initial conditions

$$\begin{bmatrix} \hat{\delta r}(t_0) \\ \hat{\delta v}(t_0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{6-15a}$$

$$\begin{bmatrix} P_{11}(t_0) & P_{12}(t_0) \\ P_{12}(t_0) & P_{22}(t_0) \end{bmatrix} = \begin{bmatrix} \sigma_r^2 & r_{rv}\sigma_r\sigma_v \\ r_{rv}\sigma_r\sigma_v & \sigma_v^2 \end{bmatrix} \tag{6-15b}$$

A block diagram of the filter is given by Fig. 6.6. The filter is seen to be a device that accepts $z(t)$, the measured difference between INS and radar (or radio) indicated positions, and outputs optimal estimates of $\delta r(t)$ and $\delta v(t)$. Note that the residual difference between $z(t)$ and $\hat{\delta r}(t)$ is put through optimal gains into a model of the system to attain the new best estimates $\hat{\delta r}(t)$ and $\hat{\delta v}(t)$.

The filter gains in (6-13) are time varying, but they converge to steady state values in a short time. By solving $\dot{\mathbf{P}}(t) = 0$, the steady state values can be shown

System model
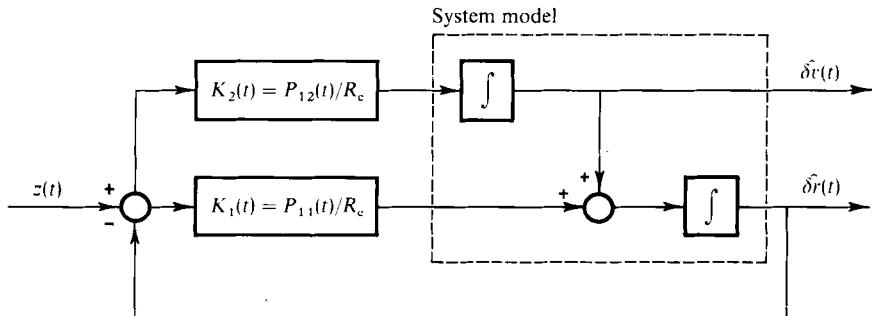
FIG. 6.6   Kalman filter block diagram.

to be covariance $\mathbf{P}$ and gain $\mathbf{K}$, where

$$\mathbf{P} = \begin{bmatrix} \sqrt{2}\,Q^{1/4}R_{\mathrm{c}}^{3/4} & Q^{1/2}R_{\mathrm{c}}^{1/2} \\ Q^{1/2}R_{\mathrm{c}}^{1/2} & \sqrt{2}\,Q^{3/4}R_{\mathrm{c}}^{1/4} \end{bmatrix} \tag{6-16a}$$

$$\mathbf{K} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} P_{11}/R_{\mathrm{c}} \\ P_{12}/R_{\mathrm{c}} \end{bmatrix} = \begin{bmatrix} \sqrt{2}\,\omega_n \\ \omega_n^{\,2} \end{bmatrix} \tag{6-16b}$$

where ω_n equals $(Q/R_{\mathrm{c}})^{1/4}$, in radians per second (the reason for denoting it as ω_n will become apparent later). The initial transient behavior of the filter gains depends on $\mathbf{P}_0$, but they are within a few percent of their steady state values (independent of $\mathbf{P}_0$) after $\omega_n t = 2$, so a prediction of time to reach steady state would be approximately $(2/\omega_n)$ sec. Note that if the noise variance kernel (or power spectral density) parameters Q and R_{c} were determined completely, so would ω_n and the gains.

This filter can be put into feedforward configuration as shown in Fig. 6.7. The optimal estimates of errors committed by the INS, $\widehat{\delta r}(t)$ and $\widehat{\delta v}(t)$, are subtracted from the INS data, to yield optimally estimated navigation information:

$$\hat{r}(t) = r_{\mathrm{i}}(t) - \widehat{\delta r}(t) \tag{6-17a}$$

$$\hat{v}(t) = v_{\mathrm{i}}(t) - \widehat{\delta v}(t) \tag{6-17b}$$

For analytical purposes to be considered presently, Fig. 6.7 shows the actual INS and radar (radio navigation aid) systems replaced by the mathematical models used to represent them.

Let us consider steady state filter operation. Under these conditions, Laplace transform techniques can be used to provide a frequency domain interpretation of the filter, and Bode amplitude ratio plots can characterize the transfer functions between INS noise $w(t)$ or radar noise $v(t)$ and the outputs $\widehat{\delta r}(t)$ or $\widehat{\delta v}(t)$.
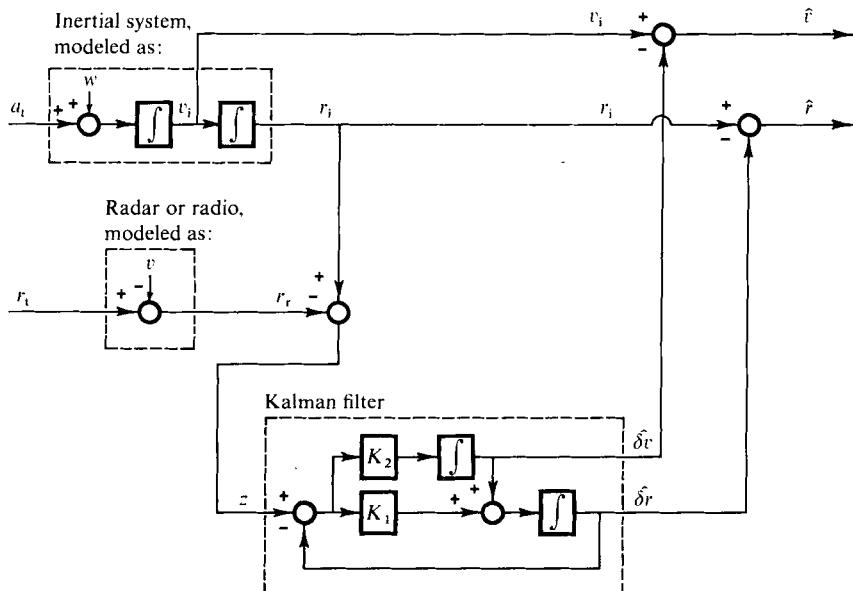
FIG. 6.7   Feedforward filter configuration.

As an intermediate step, the filter transfer function $[\hat{\delta r}(s)/z(s)]$ is given as

$$
\frac{\hat{\delta r}(s)}{z(s)} = \frac{[K_1 + (K_2/s)][1/s]}{1 + [K_1 + (K_2/s)][1/s]} = \frac{K_1[s + (K_2/K_1)]}{s^2 + K_1 s + K_2}
$$

$$
= \frac{\sqrt{2}\,\omega_n[s + (\omega_n/\sqrt{2})]}{s^2 + \sqrt{2}\,\omega_n s + \omega_n^2} \tag{6-18}
$$

Thus, the filter is a second order system with undamped natural frequency ω_n (motivating the original notation ω_n) and damping ratio $\sqrt{2}/2$: the "optimal" damping ratio that provides minimum settling time for a second order system. Furthermore, from Fig. 6.7 it is apparent that $\hat{r}(s)/r_r(s) = -\hat{r}(s)/v(s)$ and

$$
\frac{\hat{r}(s)}{v(s)} = -\frac{\hat{\delta r}(s)}{v(s)} = -\frac{\hat{\delta r}(s)}{z(s)} = \frac{-\sqrt{2}\,\omega_n[s + (\omega_n/\sqrt{2})]}{s^2 + \sqrt{2}\,\omega_n s + \omega_n^2} \tag{6-19}
$$

Thus, the transfer function from radar (radio) noise to position estimate is represented by Fig. 6.8: the filter operates as a low-pass filter on the radar, attenuating this signal at frequencies above ω_n with a 20 dB/decade rolloff, as desired. By a similar procedure, INS-caused errors are attenuated at frequencies below ω_n with a 20 dB/decade rolloff, as desired. For example, components of error at the Schuler frequency ω_s would be attenuated by the factor $(\omega_s/\omega_n)^2$.
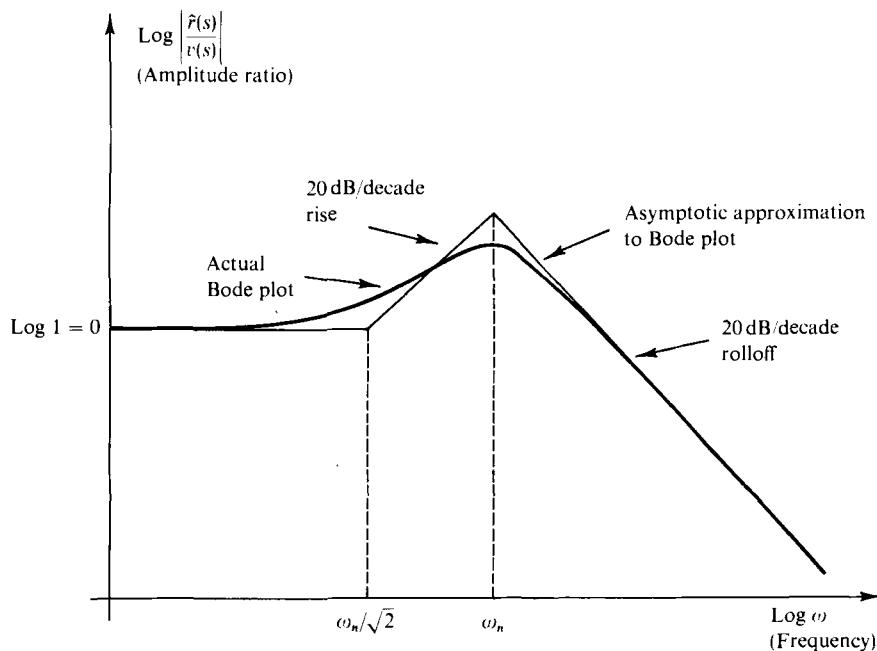
FIG. 6.8 Bode amplitude ratio plot of transfer function $[\hat{r}(s)/v(s)]$.

Figure 6.7 reveals that optimal estimates of both position and velocity are obtained from the position-type measurement *without requiring a differentiation.* This is extremely advantageous, since differentiation of a noisy signal accentuates the noise and generally yields unsatisfactory results.

Because the Kalman filter is based upon the assumed validity of a linear system model, a feedback configuration is preferable to feedforward. Perhaps the most straightforward means of generating the feedback implementation is to write the system and filter equations in terms of corrected INS states. Define

$$\hat{r}(t) = r_i(t) - \hat{\delta r}(t) \tag{6-20a}$$

$$\hat{v}(t) = v_i(t) - \hat{\delta v}(t) \tag{6-20b}$$

as the outputs of an INS corrected by feedback from the filter, and write the difference of (6-6) and (6-13) (for stochastic processes or for their realizations as done here) as:

$$
\begin{bmatrix} \dot{r}_i(t) - \hat{\dot{\delta r}}(t) \\ \dot{v}_i(t) - \hat{\dot{\delta v}}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_i(t) - \hat{\delta r}(t) \\ v_i(t) - \hat{\delta v}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [a_t(t) + w(t)]
$$
$$
- \begin{bmatrix} K_1(t) \\ K_2(t) \end{bmatrix} [z(t) - \hat{\delta r}(t)]
$$

By writing the residual as

$$[z(t) - \hat{\delta r}(t)] - [r_i(t) - \delta r(t) - r_r(t)]$$

and using (6-20), this becomes

$$\begin{bmatrix} \dot{\hat{r}}(t) \\ \dot{\hat{v}}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{r}(t) \\ \hat{v}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [a_t(t) + w(t)]$$

$$- \begin{bmatrix} K_1(t) \\ K_2(t) \end{bmatrix} [\hat{r}(t) - r_r(t)] \qquad (6\text{-}21)$$

Figure 6.9a is a block diagram representation of these equations, from which the INS, radar (radio), and filter become readily apparent. Portrayed in Fig.
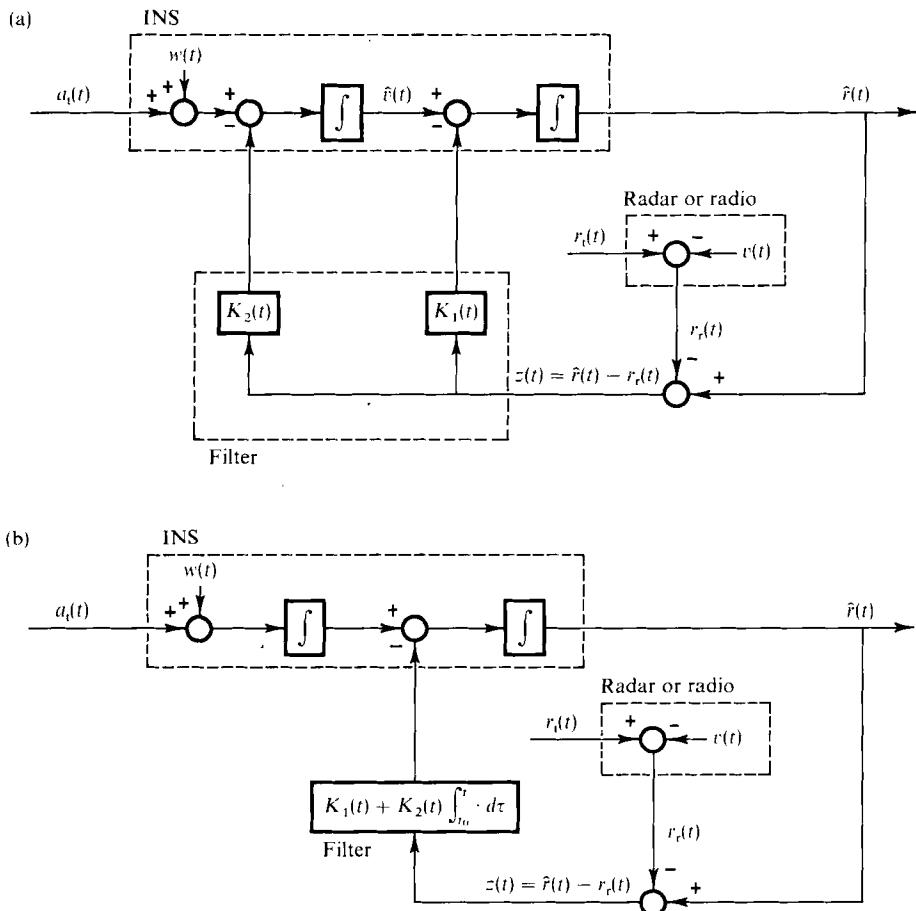
FIG. 6.9   Feedback filter configuration.

6.9b is an equivalent form obtained by replacing the feedback path through $K_2(t)$ and the first INS integrator by $K_2(t)$ and an integrator in the filter itself. Thus, with grossly approximate models for the INS and radar (radio), the filter achieved in Fig. 6.9b in steady state becomes a proportional plus integral filter with transfer function

$$G(s) = K_1 + (K_2/s) = (K_1 s + K_2)/s \qquad (6\text{-}22)$$

This is precisely the form that has been suggested by classical filtering techniques for this application. Unlike the classical approach of essentially guessing an appropriate filter transfer function form, however, optimal estimation theory *dictates* the *form* of the filter once an adequate system model has been established.

This has been an extremely simple example to demonstrate the mechanization of an optimally aided inertial navigation system and to reveal certain facets of the Kalman filter incorporated in the system integration. However, the same error state space formalism and fundamental estimation concepts are directly applied in practice to more realistic system models. By replacing the two-state INS error model with the Pinson 9-state (three position errors, three velocity errors, and three platform misalignment angles) error model [40, 41, 52], and replacing the single white noise $w(\cdot, \cdot)$ by appropriate disturbances generated from numerous shaping filters, an operational discrete-time filter can readily be developed based on the basic insights gained from this example.

A tradeoff of algorithm simplicity and estimation performance is involved in the choice of state space model: one wants to portray the dominant effects of the error dynamics well enough to attain desired accuracy for the overall navigation system, while meeting constraints of computer time, memory, wordlength, and the like.

Furthermore, once the basic model form is chosen, "best" values for noise statistics must be attained. In this problem, if Q and R_c were not known completely, engineering judgment could be employed to select values to achieve a good break frequency ω_n for the steady state filter. Iterative fine tuning of a filter will be discussed in detail in Section 6.8.

6.5 DOPPLER-AIDED INS

The functional diagram of a Doppler-aided inertial navigation system [19, 21, 25, 32, 40, 45, 52] in one axis is given in Fig. 6.10. The arrow denoted "platform orientation" is meant to indicate a physical connection rather than an electrical signal connection: what the accelerometer measures is dictated by the angular orientation of its sensitive axis, fixed to the INS platform. As is typical of terrestrial navigation systems, it will be assumed that the platform is aligned to a local-level coordinate frame (north–east–down, wander azimuth,
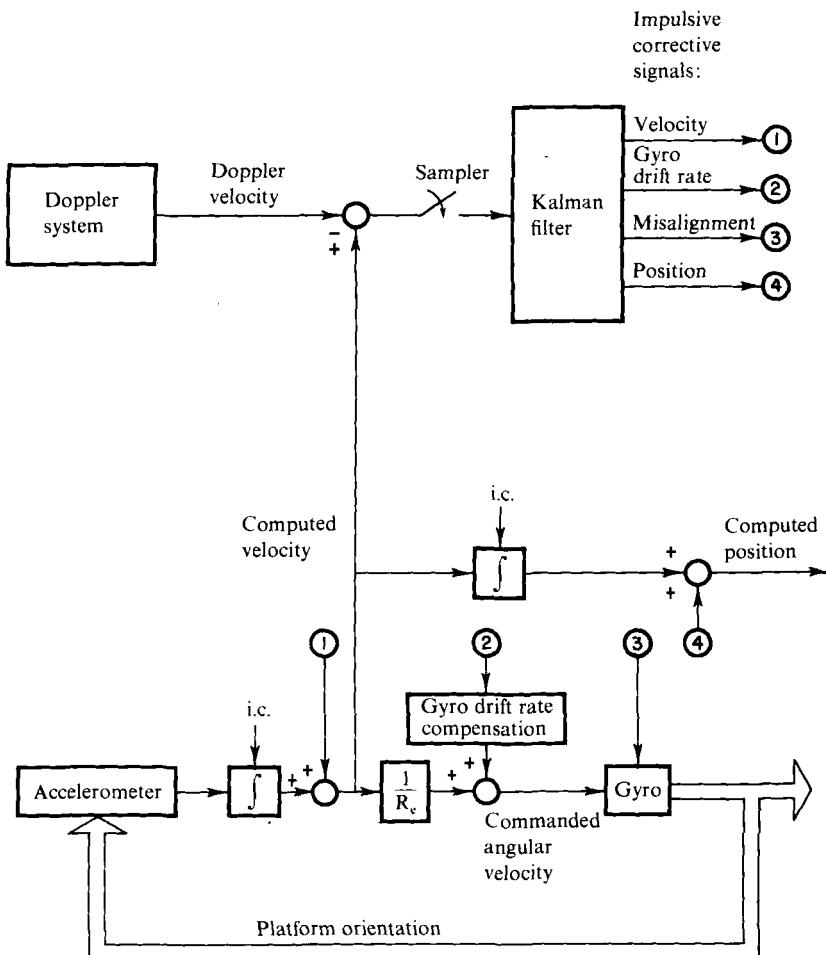
FIG. 6.10   Functional diagram of Doppler-aided INS in one axis.

etc.) and the accelerometer in Fig. 6.10 is one whose sensitive axis is nominally horizontal. (Thus, there is no computed gravity term subtracted from its output.)

The difference between INS-indicated velocity and Doppler-indicated velocity is sampled to provide discrete-time measurements to the Kalman filter, implemented in a digital computer. As can be seen from the diagram, this is an *indirect feedback* configuration: each time the measurement is sampled and $\hat{x}(t_i^+)$ is computed, the filter outputs four discrete-time corrective signals which are fed back into the INS. The four parameters to be corrected are position, velocity, platform misalignment angle, and gyro drift rate.

Position and velocity integrators (computer registers) can be corrected essentially instantaneously. Similarly, the drift rate compensation register can be altered impulsively, yielding a compensation signal (analog, or pulse rate if the INS is a pulse-torque loop design) that is held constant over the ensuing sample period. We will *assume* that instantaneous corrections to the platform misalignment are also achievable. In fact, the gimbal torquer motors are commanded to remove estimated misalignments at maximum rate, which yields a response time very short compared to the filter sample period. In some inertial systems, the outputs are resolved through a direction cosine matrix which *can* be changed instantaneously, while the gimbal motors simultaneously zero out the estimated error from the physical platform orientation, thus enhancing the validity of the assumed impulsive correction capability.

The filter is often implemented as a software program in a general purpose digital computer, inherently requiring sampled data measurements. Typically, the two velocities are sampled and differenced every 5 to 30 sec, and thus discrete-time corrections are applied to the INS with this same periodicity. For this example, we will assume a 30-sec sample period. Note that it is the use of the error state space model with its slow, linear dynamics that allows such a slow filter algorithm iteration rate to be employed.

First let us formulate the mathematical system model. Since we seek coupled first order linear differential equations, an error state space formulation will be used. Neglecting accelerometer bias (which is justifiable in many practical systems since other error sources dominate the bias effects), there are four variables of primary interest:

δr = error in INS-indicated position,
δv = error in INS-indicated velocity,
ψ = platform misalignment angle (or attitude error, tilt, or "correction to the vertical"),
ε = gyro drift rate.

In terms of these variables, the state differential equation becomes:

$$\begin{bmatrix} \dot{\delta r}(t) \\ \dot{\delta v}(t) \\ \dot{\psi}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & g & 0 \\ 0 & -1/R_e & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta r(t) \\ \delta v(t) \\ \psi(t) \\ \varepsilon(t) \end{bmatrix} \tag{6-23a}$$

$$(\dot{\mathbf{x}}(t) = \mathbf{F} \quad \mathbf{x}(t)) \tag{6-23b}$$

where g is the magnitude of the gravity vector at the earth surface and R_e is the radius of the earth (effects of vehicle altitude on g and the difference of the true geoid shape from an assumed sphere are included in navigation equations but neglected in the error state space filter). The first component of (6-23) simply says that the time rate of change of position error is equal to velocity error. To

interpret the second component, recall that a local-level inertial system is used: nominally the accelerometer sensitive axis is orthogonal to the gravitational field, but if there is a platform misalignment angle $\psi(t)$, the accelerometer senses $g \sin \psi(t)$, which for a small misalignment angle is approximately $g\psi(t)$. The time rate of change of velocity error equals the error in sensed acceleration, or $g\psi(t)$. If the vehicle were traveling at velocity $v(t)$ over an earth of radius R_e, the correct angular rate at which to command the platform to maintain it aligned to the local level would be $v(t)/R_e$, so an error in knowledge of $v(t)$ would yield an inappropriate rate command of $\delta v(t)/R_e$. The time rate of change of platform misalignment angle equals this plus gyro drift rate, with the signs in the third component of (6-23) determined by angle sign conventions. Finally, assuming gyro drift rate to be adequately modeled as an unknown constant yields the last entry in (6-23). If a random walk model of bias were used, or equivalently if a pseudonoise were to be added to reflect the fact that an unknown constant were not a totally adequate model, then (6-23) would become

$$\begin{bmatrix} \dot{\delta r}(t) \\ \dot{\delta v}(t) \\ \dot{\psi}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & g & 0 \\ 0 & -1/R_e & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta r(t) \\ \delta v(t) \\ \psi(t) \\ \varepsilon(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} w(t) \qquad (6\text{-}24)$$

where $w(\cdot,\cdot)$ is a zero-mean white Gaussian noise of appropriate strength Q. One objective of this example is to investigate the difference in filter performance caused by the two bias models.

The state initial condition is modeled as a Gaussian random variable, $\mathbf{x}(t_0)$. It is estimated that the system is error-free initially (since deterministic errors would be compensated):

$$E\{\mathbf{x}(t_0)\} = \hat{\mathbf{x}}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T \qquad (6\text{-}25)$$

The initial covariance matrix $\mathbf{P}_0$ provides a statistical measure of confidence that the states truly are error-free:

$$E\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}_0][\mathbf{x}(t_0) - \hat{\mathbf{x}}_0]^T\} = \mathbf{P}_0 = \begin{bmatrix} \sigma_{r0}^2 & 0 & 0 & 0 \\ 0 & \sigma_{v0}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\psi 0}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon 0}^2 \end{bmatrix} \qquad (6\text{-}26)$$

In this matrix, the diagonal terms represent variances, or mean squared errors, in knowledge of the initial conditions. $\mathbf{P}_0$ is often assumed diagonal for lack of sufficient statistical information to evaluate its off-diagonal terms; this infers that initially the four states are uncorrelated, or independent since $\mathbf{x}(t_0)$ is assumed to be Gaussian.

The measurement to be used as the input to the filter is the sampled difference between INS and Doppler velocities. By the definition of $\delta v(t_i)$, the INS-

indicated velocity at time t_i is modeled as

$$v_{INS}(t_i) = v_{true}(t_i) + \delta v(t_i) \tag{6-27}$$

For this problem, the Doppler velocity indication is modeled as the true velocity corrupted by the discrete-time white Gaussian noise $v(\cdot,\cdot)$, of mean zero and variance R:

$$v_{Doppler}(t_i) = v_{true}(t_i) - v(t_i) \tag{6-28}$$

Thus, the measurement can be modeled as

$$z(t_i) = v_{INS}(t_i) - v_{Doppler}(t_i) \tag{6-29}$$

$$= \delta v(t_i) + v(t_i) \tag{6-30}$$

In terms of the error state vector notation, this becomes

$$z(t_i) = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta r(t_i) \\ \delta v(t_i) \\ \psi(t_i) \\ \varepsilon(t_i) \end{bmatrix} + v(t_i) \tag{6-31a}$$

$$(z(t_i) = \qquad H \qquad x(t_i) \quad + v(t_i)) \tag{6-31b}$$

Based upon the system model of (6-23) or (6-24), (6-25), (6-26), and (6-31), the Kalman filter can now be delineated. First consider propagation from sample time t_{i-1} to time t_i. It is assumed that when the measurement is sampled, the update computations are performed and the corrective signals applied to the INS, at which point the optimal error state estimate becomes

$$\hat{x}(t_{i-1}^{+c}) = \begin{bmatrix} \hat{\delta r}(t_{i-1}^{+c}) \\ \hat{\delta v}(t_{i-1}^{+c}) \\ \hat{\psi}(t_{i-1}^{+c}) \\ \hat{\varepsilon}(t_{i-1}^{+c}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \hat{\psi}(t_{i-1}^{+}) \\ 0 \end{bmatrix}$$

where the additional superscript c denotes after the control is fed back to the inertial system. Moreover, the estimated tilt is zeroed out as quickly as possible, so that by the next sample time (and well before that time in the sample period), before the next measurement is processed, the estimated state is

$$\hat{x}(t_i^-) = \begin{bmatrix} \hat{\delta r}(t_i^-) \\ \hat{\delta v}(t_i^-) \\ \hat{\psi}(t_i^-) \\ \hat{\varepsilon}(t_i^-) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{6-32}$$

Consequently, there is no need to compute $\hat{x}(t_i^-)$ explicitly onboard.

If (6-23) is used as the state dynamics model, the covariance propagation relation is

$$\mathbf{P}(t_i^-) = \boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\boldsymbol{\Phi}^{\mathsf{T}}(t_i, t_{i-1}) \tag{6-33}$$

where $\boldsymbol{\Phi}(t_i, t_{i-1})$ is the state transition matrix associated with $\mathbf{F}$. Equivalently, $\mathbf{P}(t_i^-)$ can be found by integrating

$$\dot{\mathbf{P}}(t/t_{i-1}) = \mathbf{F}\mathbf{P}(t/t_{i-1}) + \mathbf{P}(t/t_{i-1})\mathbf{F}^{\mathsf{T}} \tag{6-34}$$

forward to time t_i from $\mathbf{P}(t_{i-1}/t_{i-1}) = \mathbf{P}(t_{i-1}^+)$ at t_{i-1}. If a random walk bias model and (6-24) were used instead, these would become, respectively,

$$\mathbf{P}(t_i^-) = \boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\boldsymbol{\Phi}^{\mathsf{T}}(t_i, t_{i-1}) + \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(t_i, \tau)\, d\tau \tag{6-35}$$

and

$$\dot{\mathbf{P}}(t/t_{i-1}) = \mathbf{F}\mathbf{P}(t/t_{i-1}) + \mathbf{P}(t/t_{i-1})\mathbf{F}^{\mathsf{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathsf{T}} \tag{6-36}$$

To update the estimate at a measurement sample time, the filter gain is calculated as

$$\mathbf{K}(t_i) = \mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}}[\mathbf{H}\mathbf{P}(t_i^-)\mathbf{H}^{\mathsf{T}} + \mathbf{R}]^{-1}$$

$$= \begin{bmatrix} K_1(t_i) \\ K_2(t_i) \\ K_3(t_i) \\ K_4(t_i) \end{bmatrix} = \frac{1}{P_{22}(t_i^-) + R} \cdot \begin{bmatrix} P_{12}(t_i^-) \\ P_{22}(t_i^-) \\ P_{32}(t_i^-) \\ P_{42}(t_i^-) \end{bmatrix} \tag{6-37}$$

Since $\hat{\mathbf{x}}(t_i^-)$ is zero, the usual state estimate update,

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[\mathbf{z}_i - \mathbf{H}\hat{\mathbf{x}}(t_i^-)]$$

simplifies to

$$\hat{\mathbf{x}}(t_i^+) = \mathbf{K}(t_i)\mathbf{z}_i = \begin{bmatrix} K_1(t_i) \\ K_2(t_i) \\ K_3(t_i) \\ K_4(t_i) \end{bmatrix} [v_{\text{INS}}(t_i) - v_{\text{Doppler}}(t_i)] \tag{6-38}$$

Finally, the covariance update is

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{K}(t_i)\mathbf{H}\mathbf{P}(t_i^-) \tag{6-39}$$

An error analysis (performance analysis) can be conducted *before* any actual hardware is built, because, under our assumptions, the estimation error covariance matrix is *not* a function of the actual measurement values. A *first step* in such an analysis would be to propagate the filter covariance equations. However, one must beware of a *misinterpretation* of these results that has been committed in some past filter design efforts. The filter-propagated error covariance is a good representation of the actual performance to be expected

only to the extent that the filter system model is an adequate representation of the "real world" environment. A more thorough and valid performance analysis will be discussed subsequently in Section 6.8.

Assume first that a filter is based upon the noise-free dynamics, (6-23), so that the filter error covariance time propagation is given by (6-33) or (6-34). Let the strength of the measurement corruption noise be

$$R = 0.25 \quad \text{ft}^2/\text{sec}^2 \tag{6-40}$$

In other words, the Doppler is modeled as a device with wideband (white) noise contributing an rms (root mean square) error of 0.5 ft/sec. Finally, let the initial state covariance be

$$\mathbf{P}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 \text{ meru}^2 \end{bmatrix} \tag{6-41}$$

where a meru is a milli-earth-rate-unit, equal to 0.015 deg/hr. This $\mathbf{P}_0$ infers that we are absolutely certain of all initial conditions except for the value of gyro drift rate. By itself, this would be rather unrealistic, but this can be viewed as one step in establishing an "error budget" to indicate overall system errors due to a *single* source of uncertainty; such a concept will be pursued further in Section 6.8.

Figure 6.11 depicts the rms errors in INS position and velocity indications and the rms platform misalignment for the case of a "pure" inertial system, running free without any Doppler aiding. These plots were generated by taking the square root of the diagonal terms of the filter-computed covariance matrix. From plot (a), it can be seen that if our model were an adequate representation of the INS, then this is about a $1\frac{1}{2}$ nautical mile per hour inertial system. Also apparent from the plots is the Schuler mode oscillation [rectified in plot (c)] with a period of 84 min, as expected from the characteristic equation associated with the $\mathbf{F}$ in (6-23):

$$|\lambda \mathbf{I} - \mathbf{F}| = 0 = \lambda^2(\lambda^2 + g/R_e) = \lambda^2(\lambda^2 + \omega_s^2) \tag{6-42}$$

Now assume that the difference between INS and Doppler velocities is sampled and processed by the filter every 30 sec. Figure 6.12 presents the corresponding rms errors in position, velocity, platform misalignment, and gyro drift rate estimates provided by the Doppler-aided inertial system. The performance improvement is impressive: the position error plot indicates achievement of an aided system with performance on the order of 1/50 nautical mile per hour. But this is a realistic prediction of true performance only if the simple four-state model adequately models true system behavior. If it were valid, one would expect about 68% of the cases of real system operation to lie within this envelope, 95% to be within an envelope of twice the magnitude,

(a)

rms position error   (nmile)

(b)

rms velocity error   (ft/sec)
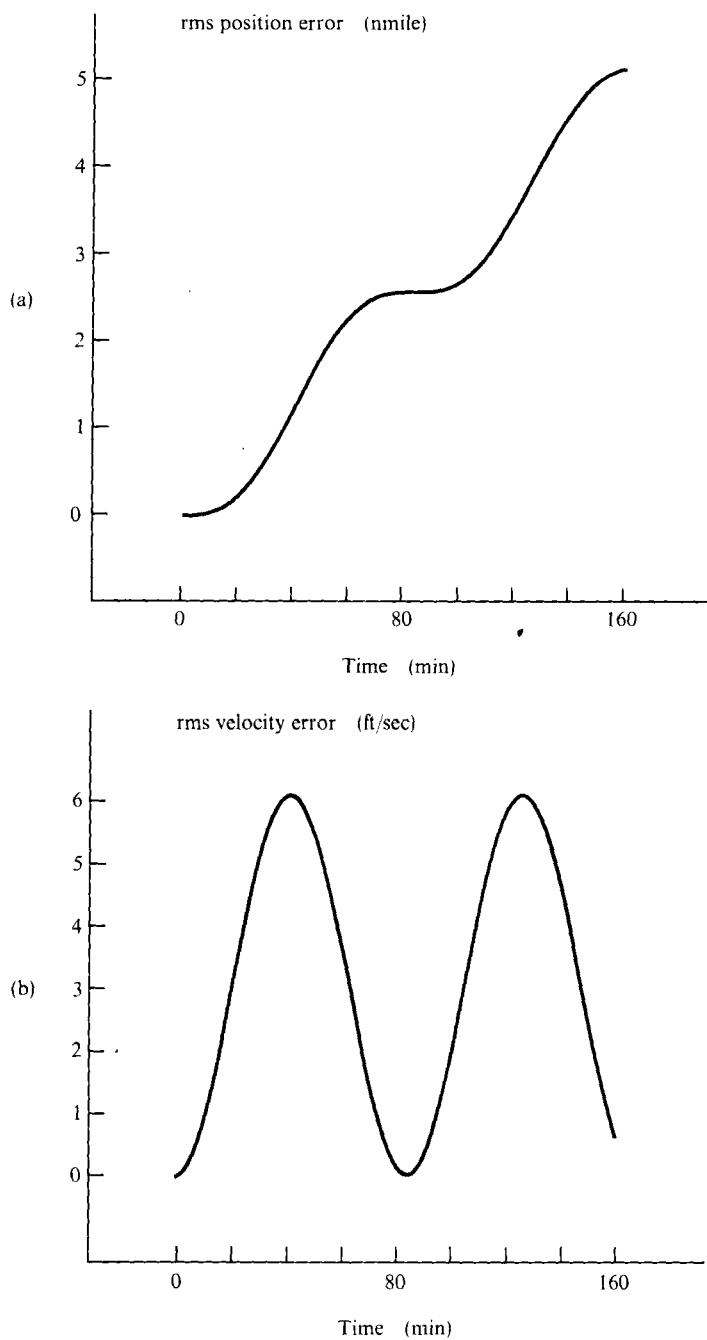
FIG 6.11a,b   Pure inertial system. (a) rms position error. (b) rms velocity error.
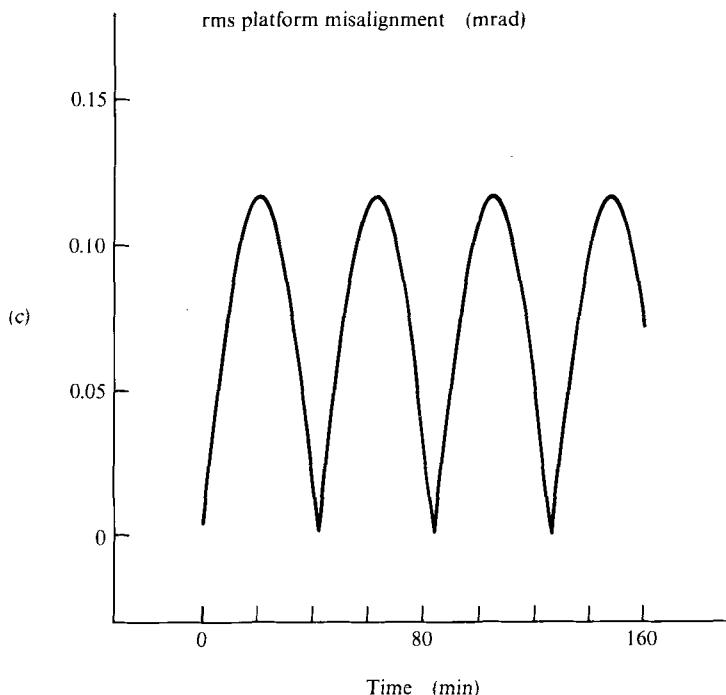
FIG. 6.11c Pure inertial system. rms platform misalignment. From Schmidt [45] with permission of author.

and so forth. Practical experience with Doppler-aided inertial systems would discount this validity significantly.

Recall that these plots pertain to the random constant bias model: there is no dynamic driving noise $w(\cdot, \cdot)$. From Fig. 6.12d it can be seen that the filter "thinks" that the error in the gyro drift rate estimate asymptotically goes to zero, and it turns out that the Kalman gain for the drift rate estimate, K_4, also goes to zero. From the postulated mathematical model, this is appropriate: with no noise $w(\cdot, \cdot)$, the model "tells" the filter that you are uncertain of the initial value of drift·rate, but that you are *sure* it is a constant value. Therefore, the filter estimates the drift rate using early measurement data, sends out the appropriate correction signals, and essentially ignores later measurements (because it "knows" once a good estimate is obtained, it is good for all time, since drift rate *is* a *constant* according to the model).

However, rarely are you so sure of drift rate (or other "biases" or parameters) being a true constant in time that you would be willing to cease estimation of its value after an initial transient period. Furthermore, this would preclude the possibility of sensor failure detection as discussed in Section 5.4. Thus, the dynamics model (6-24) is motivated, leading to filter equations as in (6-35) or
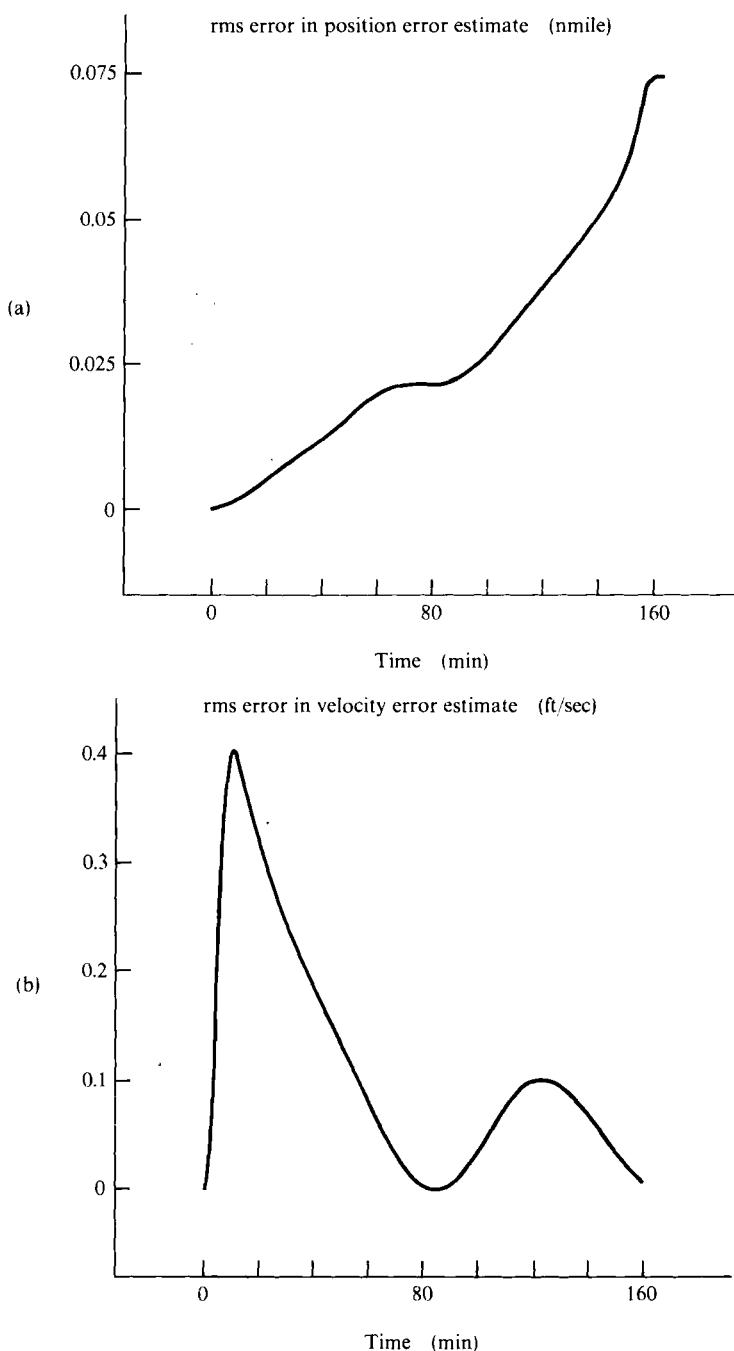
FIG. 6.12a,b   Doppler-aided inertial system (a) rms position error. (b) rms velocity error.

rms error in platform misalignment estimate   (mrad)

Time   (min)

rms error in gyro drift estimate   (meru)

Time   (min)
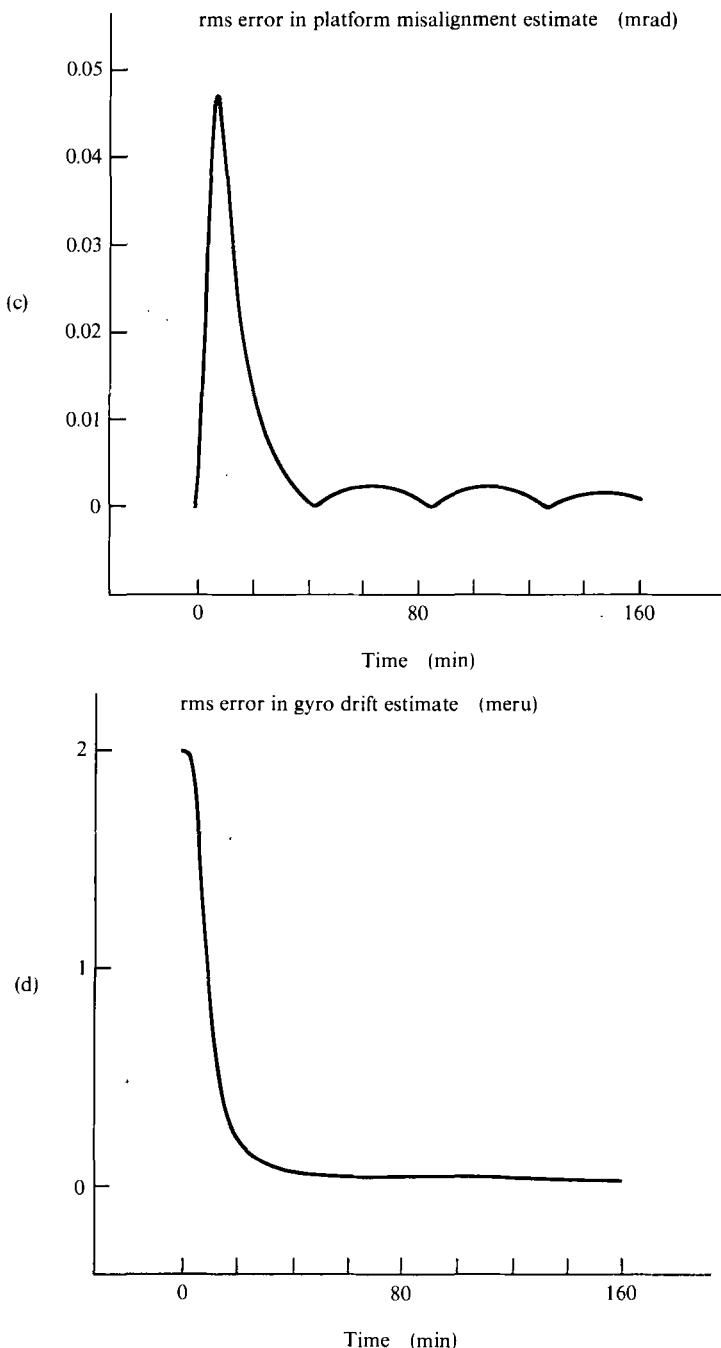
FIG. 6.12c,d   Doppler-aided inertial system. (c) rms platform misalignment. (d) rms gyro drift rate error. From Schmidt [45] with permission of author.

(6-36): a "pseudonoise" is added, of strength appropriate to how quickly you think the "bias" might actually change in time. As a result, the diagonal terms of $\mathbf{P}(t_i{}^-)$ and $\mathbf{P}(t_i{}^+)$ and the filter gains $\mathbf{K}(t_i)$ converge to nonzero values, and a valid estimate of drift rate is maintained by using measurement data over the entire time of filter operation.

One of the most effective operational Doppler/INS navigation Kalman filters (as implemented on the F-111 aircraft) is a direct extension of this example [32]. It consists of 12 states: two horizontal positions, two horizontal velocities, three platform misalignments, three gyro drift rates, and two Doppler error states. Vertical position and velocity are not maintained in the filter, but are filtered classically with barometric altimeter data; the third axis of misalignment angle, i.e., azimuth error, is modeled directly because it is critical to system performance. Essentially, the first ten states just apply the one-dimensional example to the three-dimensional case, with a slightly more complicated $\mathbf{F}(\cdot)$ to account for cross-coupling effects. The final two states are shaping filter states to allow more accurate depiction of Doppler error characteristics: like the gyro drift rate shaping filters, these are simply noise-driven integrators to generate a bias plus random walk. For use over land, they correspond to Doppler scale factor and boresight error, whereas for use over water (in which case errors caused by the water surface effects dominate) they correspond to sea motion error states in the two horizontal directions.

The filter operates in a partial feedback mode, with a sample period of 8 sec. No feedback is provided to the Doppler because the additional computation and system complexity required to do so was not warranted by performance benefits. Since Doppler errors do not exhibit the unbounded growth charac-teristic of INS errors, it is not critical for system model validity to provide such feedback. Similarly, gyro drift rate compensation feedback was removed after a tradeoff analysis indicated that the resulting hardware simplification could be gained with no appreciable performance degradation. Thus, not all components of $\hat{\mathbf{x}}(t_i{}^-)$ can be assumed to be zeros, and some state estimate time propagation computations are required.

Doppler velocity is available more often than the 8-sec filter sample period, so some smoothing or prefiltering of the velocity difference signal could enhance system performance. If one were to average the signal over an 8-sec period (with a dumped integrator), noise would in fact be smoothed out, but the resulting average would be a good representation of the signal at the *midpoint* of the integration interval, not at the end of the interval. Simply averaging the signal and inputting the result into the Kalman filter would thus introduce a 4-sec "delay" in the measurement data. To account for this directly, additional integrator states could be added to the filter model, indicating that $z(t_i)$ is not a velocity difference signal but an average (time integral over a sample period, divided by the period) of such a signal. Such an increased-dimension filter was evaluated against the 12-state design that accounted for the averaging to some

degree with a modified $\mathbf{H}(t_i)$ and $\mathbf{R}(t_i)$, and the simpler design was chosen because performance still met specifications while computer loading was decreased.

The filter also employs reasonableness checking via residual monitoring as discussed in Section 5.4. If the observed residual exceeds 2.83 times the computed standard deviation, the data point is rejected and the operator is alerted.

In early operational tests, the pure inertial mode unfortunately seemed to outperform the Doppler velocity-aided inertial mode. This was indicative of an overly pessimistic model of errors committed by the INS and/or an overly optimistic model of errors in the external velocity signal. By gathering extensive sensor performance data (unavailable at time of initial design), better values of noise strengths $\mathbf{Q}$ and $\mathbf{R}$ were established, the filter retuned, and performance improved substantially.

The point of this discussion is that the simple four-state filter discussed in this section provides the essence of a practical, proven design.

6.6 INS CALIBRATION AND ALIGNMENT USING DIRECT KALMAN FILTER

Kalman filters are exploited in the initial calibration and alignment of inertial systems as well as optimal aiding during the navigation mode of operation. A direct filter can process *external* information in order first to estimate the system misalignments and miscalibrations and then to command appropriate control signals to remove the estimated errors. The external information may simply be the fact that the vehicle is sitting still at a known location (i.e., preflight), or it may be position and/or velocity data from other sources (i.e., inflight alignment, with TACAN data, for example). In fact, experience has shown the necessity of a Kalman filter to achieving inflight alignments rapidly and precisely enough to meet most system specifications. In this section, however, the simpler problem of preflight calibration and alignment will be discussed. A rudimentary problem will be considered first, progressing in complexity to the filter form as essentially implemented for the Apollo spacecraft program.

Basically, in a preflight procedure, the accelerometers are calibrated by the known gravity acceleration magnitude, and the gyros by the known inertial rate of rotation of the gravity vector, at a point on the surface of the earth. The platform is approximately aligned to a local-level coordinate frame, and then the gyros alone are used to generate the commands appropriate to maintain an inertially fixed platform orientation. Since the estimates of platform misalignments and gyro drift rates will depend upon the measurement by the accelerometers of the rotation of the gravity vector with respect to the "inertial" frame as instrumented by the gyros, critical disturbances will be accelerometer

quantization errors and the motion of the vehicle (for instance, wind-induced sway for a missile on a launch pad). Therefore, the eventual filter design must account for these effects.

First consider a simplified single-axis problem formulation, with continuous-time measurements [42], as depicted in Fig. 6.13. The gimbal motors are driven by a known command torque, T_{com} (zero if an inertially stable platform is desired) and by gyro drift rate ε. The accelerometer sensitive axis is parallel to the platform, so if it is misaligned from local-level by an angle ψ, the accelerometer output is $g \sin \psi \cong g\psi$, corrupted by a wideband noise, modeled as white noise $v_c(\cdot, \cdot)$ of strength R_c. The purpose of the calibration filter is to accept the accelerometer data and to estimate the misalignment angle $\psi(t)$ and gyro drift rate $\varepsilon(t)$.
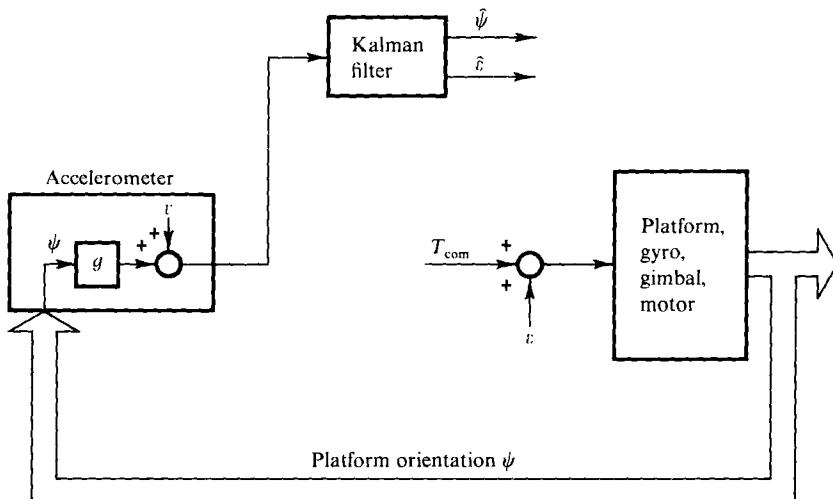
FIG. 6.13   Simplified functional diagram of INS calibration.

A very simple dynamics model would be

$$\begin{bmatrix} \dot{\psi}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \psi(t) \\ \varepsilon(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} T_{com}(t) \qquad (6\text{-}43a)$$

$$(\dot{\mathbf{x}}(t) \quad = \quad \mathbf{F} \quad \mathbf{x}(t) \quad + \quad \mathbf{B} \quad u(t)) \qquad (6\text{-}43b)$$

in which the gyro drift rate is modeled as an unknown constant. The measurement from the accelerometer can be modeled as the continuous-time relation

$$z(t) = \begin{bmatrix} g & 0 \end{bmatrix} \begin{bmatrix} \psi(t) \\ \varepsilon(t) \end{bmatrix} + v_c(t) \qquad (6\text{-}44a)$$

$$(z(t) = \quad \mathbf{H} \quad \mathbf{x}(t) \quad + v_c(t)) \qquad (6\text{-}44b)$$

Based on this model, the filter equations for calibration are

$$\mathbf{K}(t) = \mathbf{P}(t)\mathbf{H}^T\mathbf{R}_c^{-1}$$
$$= \begin{bmatrix} gP_{11}(t)/R_c \\ gP_{12}(t)/R_c \end{bmatrix} = \begin{bmatrix} K_1(t) \\ K_2(t) \end{bmatrix} \tag{6-45}$$

$$\begin{bmatrix} \dot{\hat{\psi}}(t) \\ \dot{\hat{\varepsilon}}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \hat{\psi}(t) \\ \hat{\varepsilon}(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}T_{com}(t) + \begin{bmatrix} K_1(t) \\ K_2(t) \end{bmatrix}[z(t) - g\hat{\psi}(t)] \tag{6-46a}$$

$$(\dot{\hat{\mathbf{x}}}(t) = \quad \mathbf{F} \quad \hat{\mathbf{x}}(t) \quad + \quad \mathbf{B} \quad u(t) \quad + \quad \mathbf{K}(t) \quad [z(t) - \mathbf{H}\hat{\mathbf{x}}(t)]) \tag{6-46b}$$

$$\dot{\mathbf{P}}(t) = \mathbf{F}\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T - \mathbf{P}(t)\mathbf{H}^T\mathbf{R}_c^{-1}\mathbf{H}\mathbf{P}(t) \tag{6-47}$$

integrated forward from the initial conditions $\hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0$, $\mathbf{P}(t_0) = \mathbf{P}_0$. Note that because there is no dynamic driving noise, $\mathbf{P}(t)$ will converge to $\mathbf{0}$, but it is the transient performance that is exploited in this application. The structure of the filter given by (6-46) is diagramed in Fig. 6.14a.

Now consider alignment: assume that calibration has been completed and a drift rate estimate $\hat{\varepsilon}_a$ is available, and that the platform has been torqued so as to zero out the estimated misalignment. Thus, at the time when alignment is initiated, t_a, $\hat{\varepsilon}(t_a) = \hat{\varepsilon}_a$, and $\hat{\psi}(t_a) = 0$. Now it is desired to maintain this local-level orientation, using the filter to aid in generation of the required command torque $T_{com}(t)$. Since

$$\dot{\psi}(t) = T_{com}(t) + \varepsilon(t) \tag{6-48}$$

according to our model, in order to regulate $\psi(t)$ to zero, the appropriate command would be

$$T_{com}(t) = -\hat{\psi}(t) - \hat{\varepsilon}(t) \tag{6-49}$$

Closing the feedback loop on the filter yields the result portrayed in Fig. 6.14b. This is essentially the same filter as used in calibration, except that the system itself provides the feedback of $\hat{\psi}(t)$ for a continuous Kalman filter design. The initial conditions for this filter are

$$\hat{\mathbf{x}}(t_a) = \begin{bmatrix} 0 \\ \hat{\varepsilon}_a \end{bmatrix} \tag{6-50a}$$

$$\mathbf{P}(t_a) = \begin{bmatrix} P_{11}(t_a) & P_{12}(t_a) \\ P_{12}(t_a) & P_{22}(t_a) \end{bmatrix} \tag{6-50b}$$

Note that $\mathbf{P}(t_a)$ is taken from the calibration filter and that $P_{11}(t_a)$ is *not* zero:
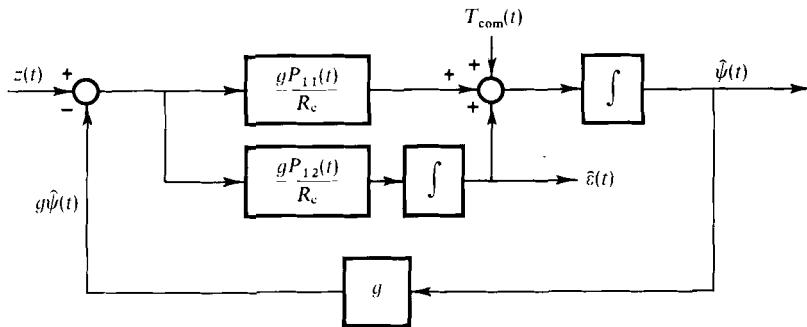
FIG. 6.14a   Simplified calibration filter.
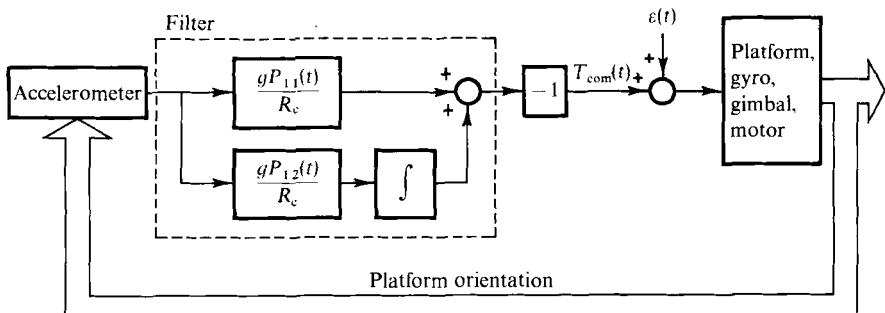
FIG. 6.14b   Simplified calibration filter.

the platform has been torqued so that $\hat{\psi}(t_a)$ is zero, but there is still uncertainty in the actual value of $\psi(t_a)$.

This simple problem can be readily modified to account for the pertinent aspects of a realistic calibration and alignment. First, the filter will be implemented on a digital computer, dictating sampled data. Moreover, the INS loop is generally a pulse-torque loop, with accelerometer signals formatted as pulse rates proportional to specific force (so each pulse is proportional to a velocity increment), and similarly torque pulses applied to the gimbal motors instead of analog signals. Through a pulse counter on the accelerometer output channel, the available discrete-time measurements are really integrals of acceleration over a sample period, corrupted mostly by the quantization of the pulses themselves. Thus, if we introduce a "dumped" integrator state (one which resets to zero after each sample time) $\delta(t)$, satisfying

$$\dot{\delta}(t) = g\psi(t) \qquad (6\text{-}51)$$

then this can be augmented to the original state equation (6-43), and the measurement model altered to the discrete-time relation

$$z(t_i) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \psi(t_i) \\ \varepsilon(t_i) \\ \delta(t_i) \end{bmatrix} + v(t_i) \tag{6-52}$$

Here $v(\cdot, \cdot)$ is a discrete-time zero-mean white Gaussian noise of strength R appropriate to model the quantization error effects.

The effects of vehicle oscillations (wind sway of a missile, roll of an aircraft carrier, etc.) can be added as well; complicated high frequency dynamics as encountered in inflight alignment would pose a more difficult problem. Consider a missile on its launch pad, and assume that acceleration of the INS case from first mode bending of the missile due to winds is significant enough to be modeled in the filter. First mode bending dynamics can be represented by a second order linear system, and a standard model of winds (the Dryden model [6]) is an exponentially time-correlated Gaussian noise. Consequently, to add this effect in one axis direction requires a three-state shaping filter; letting p_b, v_b, and a_b be the horizontal displacement, velocity, and acceleration, respectively, of the INS case due to wind-induced bending, one state space representation of the shaping filter is

$$\begin{bmatrix} \dot{p}_b(t) \\ \dot{v}_b(t) \\ \dot{a}_b(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha & \beta & \gamma \end{bmatrix} \begin{bmatrix} p_b(t) \\ v_b(t) \\ a_b(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w(t) \tag{6-53}$$

The white Gaussian noise $w(\cdot, \cdot)$ is of appropriate strength to yield the desired rms value of wind, σ_{wind}, with correlation time $1/\lambda$: $Q = 2\lambda\sigma_{\text{wind}}^2$. If the bending dynamics model is second order with undamped natural frequency ω_n and damping ratio ζ, then the three parameters α, β, and γ are specified by $\alpha = -\lambda\omega_n^2$, $\beta = -\omega_n^2 - 2\zeta\lambda\omega_n$, and $\gamma = -2\zeta\omega_n - \lambda$. Augmenting this to the dynamics given by (6-43) and (6-51) yields a six-state model with $\mathbf{x} = [\psi \ \varepsilon \ \delta \ p_b \ v_b \ a_b]^T$, and a measurement model given by

$$z(t_i) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}(t_i) + v(t_i) \tag{6-54}$$

Repeating this development for the second horizontal direction, and adding an azimuth angle error state results in the 13-state model which served as the basis of the Apollo calibration and alignment Kalman filter design. In actual implementation, the filter gains were precomputed and curve-fitted, and the approximate gain functions stored for online usage. Adequacy of this simple design is attested to by the success of the Apollo missions.

6.7 GENERATING ALTERNATIVE DESIGNS

A systematic design procedure will encompass the generation of alternative filter designs and a realistic evaluation of performance capabilities versus computer loading for each one. It is possible to devise filters based on very extensive models, but these are usually more sophisticated than really needed, and are prohibitive computationally. The designer is willing to sacrifice some degree of performance in order to achieve practical computational levels. Typically, he will seek the most simplified filter system model that retains the dominant features of the original system and provides adequate estimate precision. This is probably the most difficult aspect of designing a Kalman filter, and it requires a good understanding of the underlying physics of the system as well as competence in filtering theory.

Suppose a large-dimensioned, complex system model existed upon which a filter could be based that far exceeded performance requirements (the most complete of these to be termed a "truth model" in the next section). Since the number of multiplications (time-consuming on a computer) and additions required by the filter algorithm are proportional to n^3 and the storage is proportional to n^2 (where n is the dimension of the state vector), one significant means of reducing the computer burden is to *delete and combine states* [31, 33, 46, 47, 49]. There is often substantial physical insight into the relative significance of various states upon overall estimation precision that suggests which states might be removed. States with consistently small rms value especially warrant inspection for possible removal. An error budget performance analysis of the most complete filter, to be discussed in the next section, is an invaluable aid to this state dimension reduction.

EXAMPLE 6.1   A standard error model of an accelerometer [2, 4, 21, 41, 52] is as follows. Let an orthogonal coordinate system be denoted as having axes x, y, and z. Then the error in the output of the accelerometer whose sensitive axis is along the x coordinate direction, α_x, is:

$$\alpha_x = a_{bx} + [f_x]\,\delta k_x + [f_y]\,\delta k_{xy} + [f_z]\,\delta k_{xz} + a_{rbx} + a_{rbx2}$$

where a_{bx} is bias error (modeled as a random constant plus random walk), δk_x is scale factor error (modeled as a random constant), δk_{xy} is sensor misalignment about the y axis (modeled as a random constant), δk_{xz} is sensor misalignment about the z axis (modeled as a random constant), a_{rbx} is "random bias" error (modeled as an exponentially time-correlated noise, with short correlation time on the order of minutes), a_{rbx2} is "random bias" error (modeled as an exponentially time-correlated noise, with long correlation time, typically an hour or longer), and where f_x, f_y, and f_z are the components of true specific force relative to the x–y–z coordinate frame.

In most operational aided INS Kalman filters, only a single bias state is included, if any at all. A white noise source is added to indicate both the accelerometer error effects and the additional model uncertainty due to the deleted states.   ■

EXAMPLE 6.2   With respect to the same x–y–z coordinates of Example 6.1, a generic model for the errors in a single-degree-of-freedom gyro with sensitive axis along the x coordinate direction is [2, 4, 21, 41, 52]:

$$e_x = \varepsilon_{dx} + [\omega_{icx}]\,\delta k_x + \{[\omega_{icy}]\,\delta k_{xy} + [\omega_{icz}]\,\delta k_{xz}\} + \omega_{rbx} + \{[f_x]k_{xx} + [f_y]k_{xy} + [f_z]k_{xz}\}$$
$$+ \{f_x^2\}k_{xxx} + \{f_y^2\}k_{xyy} + \{f_z^2\}k_{xzz}\} + \{[f_x f_y]k_{xxy} + [f_x f_z]k_{xxz} + [f_y f_z]k_{xyz}\} + w_x$$

where ε_{dx} is gyro drift rate bias error (modeled as a random constant plus random walk), δk_x is scale factor error (modeled as a random constant), δk_{xi} is sensor misalignment about the $i = y$ or z axis (modeled as a random constant), ω_{rbx} is time-correlated gyro drift rate (modeled as an exponentially time-correlated noise), k_{xi} is a "g-sensitive error," sensitive to specific force in the $i = x, y,$ or z direction (modeled as a random constant), k_{xii} is a "g^2-sensitive error," sensitive to the square of specific force in the $i = x, y,$ or z direction (modeled as a random constant), k_{xij} is a "cross-term g^2-sensitive error," sensitive to the product of specific forces in the i and j directions, $i \neq j$ (modeled as a random constant), and w_x is the white noise gyro drift rate (not requiring a state). The coefficients ω_{icx}, ω_{icy}, and ω_{icz} are the components of the angular rate between inertial space and the x–y–z coordinate system, as measured in that system.

For conventional (fluid- or dry-tuned) gyros, only the ε_{dx} or ω_{rbx} state is usually retained for aided INS Kalman filters, since this drift rate predominates. In laser gyros, the w_x noise predominates (and the g- and g^2-sensitive errors are essentially zero), so no states are required in the Kalman filter for gyro errors, only a white noise model.   ■

EXAMPLE 6.3   One conventional approach to reducing state dimension is to attempt to replace exponentially time-correlated noises (requiring single-state shaping filters) with appropriate white noises (requiring no states). Consider a stationary exponentially time-correlated zero-mean noise with rms value σ and correlation time T, i.e., with autocorrelation

$$\Psi_{xx}(\tau) = E\{x(t)x(t + \tau)\} = \sigma^2 e^{-|\tau|/T}$$

with $w(\cdot, \cdot)$ a zero-mean white Gaussian noise of strength $Q = 2\sigma^2/T$. The power spectral density for $x(\cdot, \cdot)$ is

$$\Psi_{xx}(\omega) = \frac{2\sigma^2/T}{\omega^2 + (1/T)^2}$$

which is plotted in Fig. 6.15.

The adequacy of replacing this noise by a white Gaussian noise is based on the premise that the system driven by the noise will attenuate high frequency content. Therefore, an appropriate strength to choose for the white noise is that which duplicates the low frequency power spectral density of $x(\cdot, \cdot)$, namely $2\sigma^2 T$.   ■
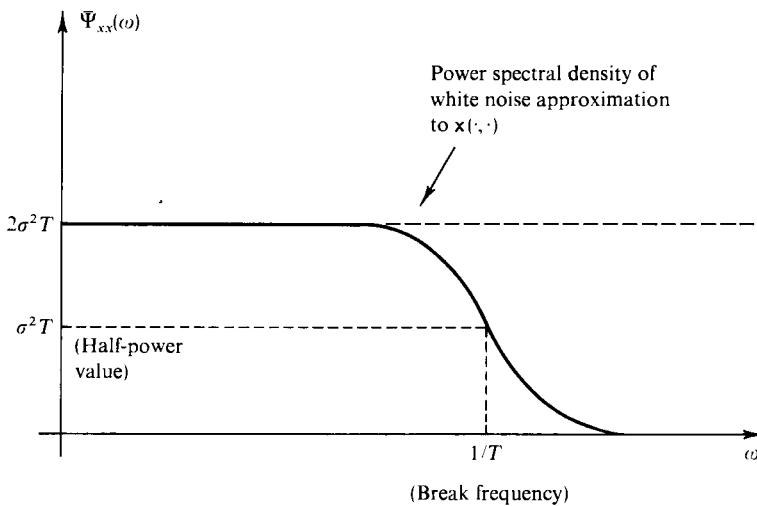
FIG. 6.15   White noise approximation to exponentially time-correlated noise.

It must be emphasized that deleting states and combining many states into fewer "equivalent" states must be evaluated in terms of resulting filter performance, as described in the next section. Experience has shown that reductions motivated by the best of physical insight can sometimes degrade estimation accuracy unacceptably. Moreover, an inappropriately reduced filter of state dimension n can often be outperformed by a filter involving less than n, differently chosen, states. The extreme case of this would be the higher-dimensioned filter being based on an unobservable model, for instance, in which two states correspond to different physical variables but which are indistinguishable from outputs from the model, while the lower-dimensioned filter model combined states to achieve observability.

The number of additions and multiplications required by a filter algorithm can be minimized by exploiting *canonical state variables*, since the matrices in this equivalent representation embody a high density of zeros. There is also some computational advantage to decomposing vector measurements into separate scalars or partitions, and *iteratively updating with lower-dimensioned measurements*. Obviously, one might also attempt to reduce the computational burden by *increasing the sample period* of the filter.

Once a filter dimension and structure are established, it is often possible to *neglect dominated terms* within the matrix elements. Moreover, entire *weak coupling terms can be removed*, yielding matrix entries of zero and thereby decreasing the number of required multiplications. In certain applications, such removal allows *decoupling the filter states* to generate separate, smaller filters.

EXAMPLE 6.4   In the Pinson error model for an INS implementing a north–east–down platform coordinate frame, one term in the $\mathbf{F}(t)$ matrix is $[2\omega_{ie}R_e \cos L(t) - v_e(t)]/R_e$, in which ω_{ie} is the earth rotation rate, $L(t)$ is current latitude, $v_e(t)$ is eastward vehicle velocity, and R_e is the earth radius. Since $[\omega_{ie}R_e]$ is on the order of 1000 knots, $v_e(t)$ can be neglected except for very high speed aircraft or near-polar operations. This entire term is small compared to Schuler frequency-dominated terms; when these weak coupling terms are ignored, the entire filter for aiding the INS with external source data often decouples into a horizontal plane filter and a separate vertical channel filter, with an acceptable (minute) amount of performance degradation.   ∎

Sometimes terms that can be ignored comprise the time-varying nature of the model description (or at least the most rapid variations), so that the filter can be based upon a *time-invariant model* (or at least a model that admits quasi-static analysis). This inherently yields computational advantages such as a single value for $\mathbf{\Phi}$, $\mathbf{B}_d$, and $\mathbf{Q}_d$ being valid for all time.

The methods discussed to this point have involved the generation of a simplified model, with subsequent filter construction. It is also advantageous to consider *approximating the filter structure* itself. Because of the separability of the conditional mean and covariance equations in the filter, it is possible to precompute and store the filter gains rather than calculate them online. This *precomputed filter gain history can often be approximated closely by curve-fitted*
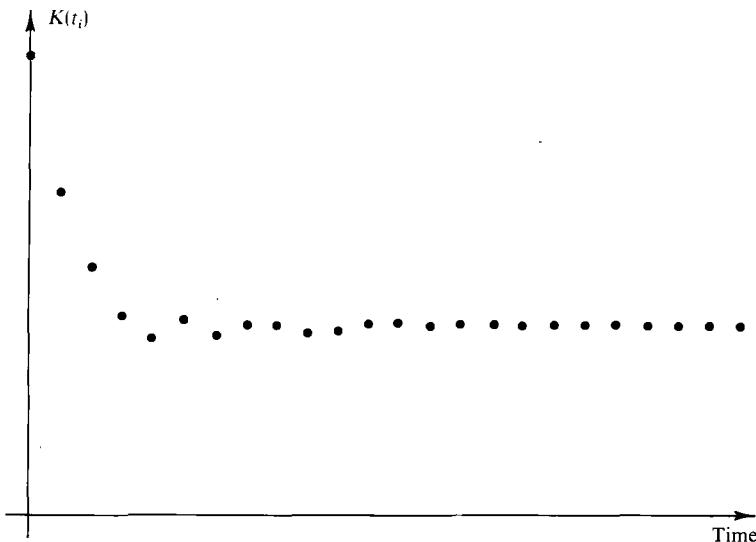
FIG. 6.16   Characteristic gain time history.

simple functions, such as piecewise constant functions, piecewise linear functions, and weighted exponentials. Thus, the filter covariance and gain calculations, which comprise the majority of the computer burden, are replaced by a minimal amount of required computation and storage. This is a tremendous benefit to the practicality of the online filter operation. For the case of a filter based on a time-invariant system model with stationary noises, a typical gain time history is depicted in Fig. 6.16. If, as in this plot, a short initial transient is followed by a long period of essentially steady state gain, the very simple approximation of using steady state gains for all time may be wholly adequate for desired performance. There are some drawbacks to using stored gain profiles. Future gains do not change appropriately when scheduled measurements are not made, due to data gaps or measurement rejection by reasonableness tests. Nor can prestored gains adapt online to compensate for filter divergence. Finally, lengthy simulations are often required to design a single gain history that will perform adequately under all possible conditions for an actual application.

6.8   PERFORMANCE (SENSITIVITY) ANALYSIS

Throughout the previous sections, the critical significance of an "adequate" system model within the filter structure was stressed. The question remains, how do you assess the adequacy of various filter designs relative to each other and/or to a set of performance specifications?
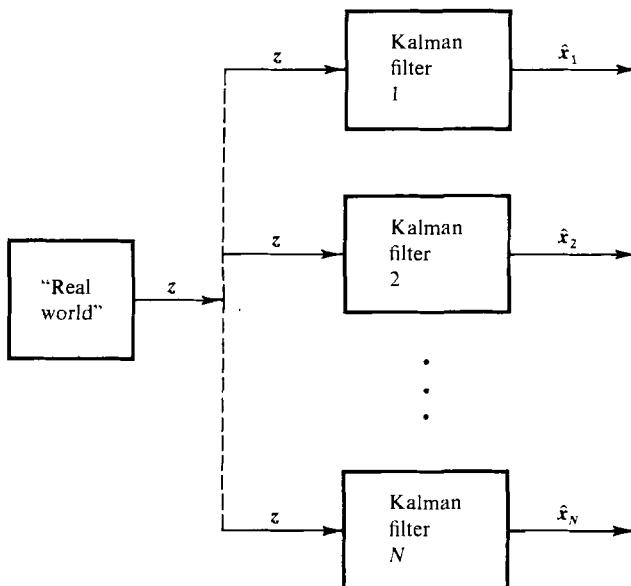
FIG. 6.17 The context of performance (sensitivity) analyses.

Consider Fig. 6.17, a schematic presentation of the situation under study. Suppose that system modeling and filter design efforts have produced N different prospective Kalman filters. Each is based upon a particular model of the "real world": a linear system with structure defined by $\{F, \mathbf{B}, \mathbf{G}, \mathbf{H}\}$ (or $\{\Phi, \mathbf{B_d}, \mathbf{G_d}, \mathbf{H}\}$ in terms of an equivalent discrete-time model) and uncertainties defined by $\{\hat{\mathbf{x}}_0, \mathbf{P}_0, \mathbf{Q}, \mathbf{R}\}$ (or $\{\hat{\mathbf{x}}_0, \mathbf{P}_0, \mathbf{Q_d}, \mathbf{R}\}$). The filters can vary greatly due to model differences, over a spectrum of small state dimension and deliberately crude approximations to high-dimensioned filters incorporating more system modes, cross-coupling effects, and high-order shaping filters for accurate portrayal of stochastic process characteristics. They might also differ due to aspects of algorithm implementation, such as Joseph versus standard covariance measurement update, gain history approximations, wordlength variations, etc.

In actual operation, any one of these filters would be driven by sampled data measurements from actual sensors operating in the "real world" environment. To make rational design decisions, one must have at his disposal an *accurate statistical portrayal of estimation errors committed by each filter in the* *"real world" environment.* Moreover, this information must be generated without actually building each filter and testing it in the "real world." A performance (sensitivity) analysis fulfills this objective by replacing the "real world" in Fig. 6.17 by the best, most complete mathematical model that can be developed, called a *truth model* or "reference model." Such a truth model is the product of extensive data analysis, and shaping filter design and validation, as

described in Section 4.13. It is essential to expend enough effort in its generation to be confident that *it* adequately represents the "real world," since the ensuing performance evaluation and systematic design procedure is totally dependent upon this assumption. For example, a very adequate generic model of the errors in an inertial navigation system has been constructed over the years in the form of a linear model of about 70 states driven by white Gaussian noise; thorough laboratory and flight testing of a particular INS allows complete specification of the model parameters to yield the "truth model" for that system. As inertial systems have improved, the "truth model" itself has become more refined so as to portray system characteristics accurately that were once considered insignificant. For example, in the next generation of systems, a dominant error source in addition to accelerometer and gyro uncertainties will be the difference between the true earth shape (geoid) and the assumed ellipsoid used in navigation computations. Without incorporating this effect, a "truth model" would be seriously inadequate.

It is desired to achieve a direct comparison of performance capabilities of filters that may well estimate completely different, and different numbers of, state variables. However, for any given application, there are certain variables of critical interest. In optimally aided inertial systems, for example, the nine variables of position, velocity, and attitude indications in the three axis directions are paramount. All prospective filters will estimate these quantities or variables functionally related to them. These critical variables, which we will denote as $y(\cdot, \cdot)$, will serve as the basis of comparison of the filter designs. Although a *scalar* performance index is appealing from a mathematical and numerical optimization point of view, in practice the designer really seeks information about these critical variables individually, so attention will be focused on them.

Figure 6.18 depicts the means of conducting a performance analysis of a given Kalman filter design [8, 31, 33, 34, 36]. The truth model is an n_t-dimensional
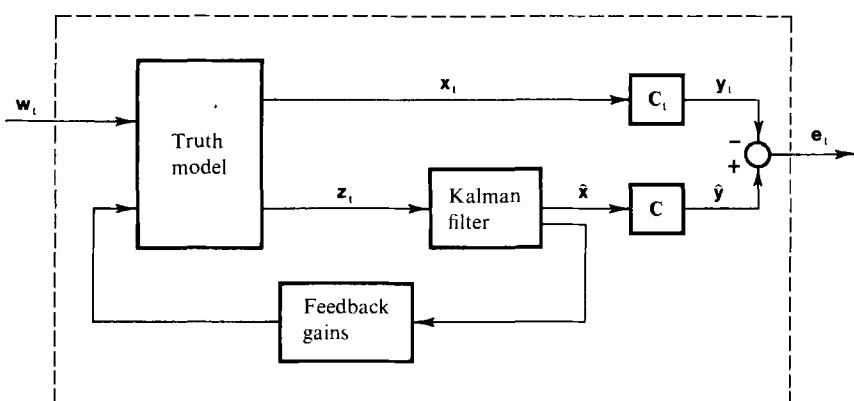
FIG. 6.18 Performance evaluation of a Kalman filter design.

state model, linear or nonlinear, driven by (white Gaussian) noise $\mathbf{w}_t(\cdot,\cdot)$, that accurately generates the measurement process $\mathbf{z}_t(\cdot,\cdot)$. The subscript t will be used to denote quantities and processes associated with the truth model. A sample from this discrete-time process, $\mathbf{z}_t(t_i,\omega_j) = \mathbf{z}_t(t_i)$ for all time $t_i \in T$, is then a representation of a measurement time history to be used as input for a single run of the filter algorithm. The filter operates on this input and generates the state estimate process $\hat{\mathbf{x}}(\cdot,\cdot)$. If it is implemented in feedback fashion as described in the preceding sections, the filter can also output feedback controls to the actual system, so a possible feedback path from the filter to the truth model is shown in the figure. Suppose that the p critical quantities are related to the filter states through a *linear* transformation:

$$\hat{\mathbf{y}}(t_i^-) = \mathbf{C}(t_i)\hat{\mathbf{x}}(t_i^-) \tag{6-55a}$$

$$\hat{\mathbf{y}}(t_i^+) = \mathbf{C}(t_i)\hat{\mathbf{x}}(t_i^+) \tag{6-55b}$$

Usually $\mathbf{C}(t_i)$ is a time-invariant p-by-n matrix, and often with partitions as a p-by-p identity matrix and p-by-$(n - p)$ zero matrix; i.e., the critical quantities compose the first p components of the filter state. Nonlinear functional relationships would be handled in an analogous manner, but we emphasize use of linear models here. Thus, the optimal estimates of the quantities of interest are generated as $\hat{\mathbf{y}}(t_i^-,\cdot)$ and $\hat{\mathbf{y}}(t_i^+,\cdot)$ for all times $t_i \in T$.

Being a mathematical representation, the truth model, unlike the "real world," generates a (continuous-time) state process $\mathbf{x}_t(\cdot,\cdot)$ as well as a measurement process $\mathbf{z}_t(\cdot,\cdot)$. We will assume that the true values of the critical quantities are related to this state by a linear transformation (again easily extended to a nonlinear function) represented by a p-by-n_t matrix $\mathbf{C}_t$ (often time-invariant):

$$\mathbf{y}_t(t,\cdot) = \mathbf{C}_t(t)\mathbf{x}_t(t,\cdot) \tag{6-56}$$

for all $t \in T$. This affords access to these true values, something which the "real world" denies us. Using (6-55) and (6-56), we can generate the "true" error committed by the Kalman filter in attempting to estimate the quantities of interest at time t_i, before and after measurement incorporation:

$$\mathbf{e}_t(t_i^-,\cdot) = \hat{\mathbf{y}}(t_i^-,\cdot) - \mathbf{y}_t(t_i,\cdot) \tag{6-57a}$$

$$\mathbf{e}_t(t_i^+,\cdot) = \hat{\mathbf{y}}(t_i^+,\cdot) - \mathbf{y}_t(t_i,\cdot) \tag{6-57b}$$

If we admit impulsive system response to the feedback control from the filter, as discussed in Section 6.5, then a third error, corresponding to after both measurement incorporation and impulsive control application, is of interest as well:

$$\mathbf{e}_t(t_i^{+c},\cdot) = \hat{\mathbf{y}}(t_i^{+c},\cdot) - \mathbf{y}_t(t_i^c,\cdot) \tag{6-57c}$$

The superscript c denotes after control is applied.

The objective of the performance analysis is to characterize the error process (6-57) statistically. It is also called a sensitivity analysis because we wish to evaluate the sensitivity of this performance to changing the filter structure. In addition, we may well want to study the effects of changing sensor hardware or system environment; i.e., altering the truth model. Because of using stochastic processes as the basic modeling entity, we are more interested in the statistical, or ensemble average, behavior of the error process than in a single sample out of this process.

One means of generating this statistical information is a *Monte Carlo study*. Essentially, many samples of the error process are generated by simulation and then the sample statistics computed directly. If enough samples are generated, these should approximate the process statistics very well. Unfortunately, this is a costly and time-consuming process.

If the truth model itself is in the form of a *linear* system driven by white Gaussian noise, from which are available *linear* measurements corrupted by white Gaussian noise, there is another, more efficient means of generating the statistical information, namely, a *mean analysis* [14] and a *covariance analysis* [11, 17, 18, 31]. Especially in the case of using an error state space Kalman filter, the means of all processes are often assumed to be zero for all time, and attention is concentrated on the covariance analysis. From *one* run of a covariance analysis is generated the time history of $\mathbf{P}_e(\cdot)$, the covariance of the true estimation errors committed by a given filter; the square roots of its diagonal terms yield the time histories of standard deviations (or "one-sigma values," equal to rms values if processes are zero mean) of errors in the estimates of the critical quantities of interest. This is directly comparable to the massive task of running the filter repeatedly,. storing all pertinent performance data, and computing sample statistics, as required for a Monte Carlo analysis of the same filter.

To develop the performance analysis equations, consider Fig. 6.18 again. If the truth model is itself a linear system model, then the entire system enclosed by the dashed lines is nothing but a linear system driven only by white Gaussian noise. As seen in Chapter 4, if we want to characterize the output process from such a system model, we must first characterize the state process within the system, in this case being composed of the partitions $\mathbf{x}_t(\cdot, \cdot)$ of the truth model and $\hat{\mathbf{x}}(\cdot, \cdot)$ of the filter under investigation. First we will develop the stochastic process description appropriate to a Monte Carlo analysis, and from it develop the statistical description appropriate to a covariance analysis. The derivation will assume all processes to be zero mean, appropriate to error state space Kalman filters, and therefore deterministic inputs $\mathbf{u}(\cdot)$ are neglected; an extension relaxing these assumptions would be straightforward but more complicated to account for nonzero means and biased estimates.

The *truth model* is described by the stochastic differential n_t-dimensional state equation

$$\mathbf{dx}_t(t) = \mathbf{F}_t(t)\mathbf{x}_t(t)\,dt + \mathbf{G}_t(t)\mathbf{d\beta}_t(t) \qquad (6\text{-}58a)$$

or

$$\dot{\mathbf{x}}_t(t) = \mathbf{F}_t(t)\mathbf{x}_t(t) + \mathbf{G}_t(t)\mathbf{w}_t(t) \tag{6-58b}$$

where $\boldsymbol{\beta}_t(\cdot, \cdot)$ is an s_t-vector Brownian motion of diffusion $\mathbf{Q}_t(t)$ for all $t \in T$:

$$E\{\boldsymbol{\beta}_t(t)\} = \mathbf{0} \tag{6-59a}$$

$$E\{[\boldsymbol{\beta}_t(t) - \boldsymbol{\beta}_t(t')][\boldsymbol{\beta}_t(t) - \boldsymbol{\beta}_t(t')]^{\mathrm{T}}\} = \int_{t'}^{t} \mathbf{Q}_t(\tau) \, d\tau \tag{6-59b}$$

or $\mathbf{w}_t(\cdot, \cdot)$ is an s_t-vector zero-mean white Gaussian noise of strength $\mathbf{Q}_t(t)$ for all $t \in T$:

$$E\{\mathbf{w}_t(t)\} = \mathbf{0} \tag{6-60a}$$

$$E\{\mathbf{w}_t(t)\mathbf{w}_t^{\mathrm{T}}(t')\} = \mathbf{Q}_t(t)\delta(t - t') \tag{6-60b}$$

Note that (6-58) does not as yet account for feedback from the Kalman filter: the necessary modifications will be incorporated once the filter is described. The initial condition for this differential equation is that $\mathbf{x}_t(t_0)$ is described as a zero-mean Gaussian random variable with covariance $\mathbf{P}_{t0}$:

$$E\{\mathbf{x}_t(t_0)\} = \mathbf{0} \tag{6-61a}$$

$$E\{\mathbf{x}_t(t_0)\mathbf{x}_t^{\mathrm{T}}(t_0)\} = \mathbf{P}_{t0} \tag{6-61b}$$

Available from the truth model at discrete times t_i are the m-dimensional measurements $\mathbf{z}_t(t_i, \cdot)$:

$$\mathbf{z}_t(t_i) = \mathbf{H}_t(t_i)\mathbf{x}_t(t_i) + \mathbf{v}_t(t_i) \tag{6-62}$$

where $\mathbf{v}_t(\cdot, \cdot)$ is a discrete-time m-vector white Gaussian noise described by

$$E\{\mathbf{v}_t(t_i)\} = \mathbf{0} \tag{6-63a}$$

$$E\{\mathbf{v}_t(t_i)\mathbf{v}_t^{\mathrm{T}}(t_j)\} = \begin{cases} \mathbf{R}_t(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{6-63b}$$

In terms of the truth model state, the quantities of direct interest to the performance evaluation are denoted as the continuous-time p-vector process $\mathbf{y}_t(\cdot, \cdot)$:

$$\mathbf{y}_t(t) = \mathbf{C}_t(t)\mathbf{x}_t(t) \tag{6-64}$$

The Kalman filter being analyzed is based upon a different, generally lower-dimensioned, system model, called a *design model* (this model never explicitly appears in the performance analysis, it just serves to generate the filter). The n-dimensional design state equation is

$$\mathbf{dx}(t) = \mathbf{F}(t)\mathbf{x}(t) \, dt + \mathbf{G}(t) \, \mathbf{d\beta}(t) \tag{6-65a}$$

or

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{w}(t) \tag{6-65b}$$

with s-vector driving noise of diffusion (strength) $\mathbf{Q}(t)$ for all $t \in T$:

$$E\{\boldsymbol{\beta}(t)\} = \mathbf{0} \tag{6-66a}$$

$$E\{[\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t')][\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t')]^{\mathrm{T}}\} = \int_{t'}^{t} \mathbf{Q}(\tau)\, d\tau \tag{6-66b}$$

$$E\{\mathbf{w}(t)\} = \mathbf{0} \tag{6-67a}$$

$$E\{\mathbf{w}(t)\mathbf{w}^{\mathrm{T}}(t')\} = \mathbf{Q}(t)\delta(t - t') \tag{6-67b}$$

The initial condition $\mathbf{x}(t_0)$ is Gaussian with

$$E\{\mathbf{x}(t_0)\} = \mathbf{0} \tag{6-68a}$$

$$E\{\mathbf{x}(t_0)\mathbf{x}^{\mathrm{T}}(t_0)\} = \mathbf{P}_0 \tag{6-68b}$$

The design model of the measurement history is the discrete-time m-vector process $\mathbf{z}(\cdot, \cdot)$

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \tag{6-69}$$

with $\mathbf{v}(\cdot, \cdot)$ a discrete-time m-vector white Gaussian noise with

$$E\{\mathbf{v}(t_i)\} = \mathbf{0} \tag{6-70a}$$

$$E\{\mathbf{v}(t_i)\mathbf{v}^{\mathrm{T}}(t_j)\} = \begin{cases} \mathbf{R}(t_i) & t_i = t_j \\ \mathbf{0} & t_i \neq t_j \end{cases} \tag{6-70b}$$

Finally, the outputs of most concern are described by the p-vector $\mathbf{y}(\cdot, \cdot)$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) \tag{6-71}$$

Despite the notational similarities of (6-65)–(6-71) and (6-58)–(6-64), the two models are distinctly different.

The *Kalman filter* based upon this design model is specified between update times by the time propagation equations

$$\hat{\mathbf{x}}(t_i^-) = \boldsymbol{\Phi}(t_i, t_{i-1})\hat{\mathbf{x}}(t_{i-1}^+) \tag{6-72a}$$

$$\mathbf{P}(t_i^-) = \boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\boldsymbol{\Phi}^{\mathrm{T}}(t_i, t_{i-1}) + \mathbf{Q}_{\mathrm{d}}(t_{i-1}) \tag{6-72b}$$

$$\mathbf{Q}_{\mathrm{d}}(t_{i-1}) = \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}(t_i, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\boldsymbol{\Phi}^{\mathrm{T}}(t_i, \tau)\, d\tau \tag{6-72c}$$

or, equivalently,

$$\dot{\hat{\mathbf{x}}}(t/t_{i-1}) = \mathbf{F}(t)\hat{\mathbf{x}}(t/t_{i-1}) \tag{6-73a}$$

$$\dot{\mathbf{P}}(t/t_{i-1}) = \mathbf{F}(t)\mathbf{P}(t/t_{i-1}) + \mathbf{P}(t/t_{i-1})\mathbf{F}^{\mathrm{T}}(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) \tag{6-73b}$$

solved forward to time t_i from the initial conditions

$$\hat{\mathbf{x}}(t_{i-1}/t_{i-1}) = \hat{\mathbf{x}}(t_{i-1}^+) \tag{6-73c}$$

$$\mathbf{P}(t_{i-1}/t_{i-1}) = \mathbf{P}(t_{i-1}^+) \tag{6-73d}$$

At measurement sample times, the update relations are

$$\mathbf{K}(t_i) = \mathbf{P}(t_i{}^-)\mathbf{H}^{\mathrm{T}}(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i{}^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i)]^{-1} \tag{6-74a}$$

$$\hat{\mathbf{x}}(t_i{}^+) = \hat{\mathbf{x}}(t_i{}^-) + \mathbf{K}(t_i)[\mathbf{z}_t(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i{}^-)] \tag{6-74b}$$

$$\mathbf{P}(t_i{}^+) = \mathbf{P}(t_i{}^-) - \mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i{}^-) \tag{6-74c}$$

$$= [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\mathbf{P}(t_i{}^-)[\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]^{\mathrm{T}} + \mathbf{K}(t_i)\mathbf{R}(t_i)\mathbf{K}^{\mathrm{T}}(t_i) \tag{6-74d}$$

Note the appearance of $\mathbf{z}_t(\cdot,\cdot)$ in (6-74b). Initial conditions for the algorithm are

$$\hat{\mathbf{x}}(t_0) = \mathbf{0} \tag{6-75a}$$

$$\mathbf{P}(t_0) = \mathbf{P}_0 \tag{6-75b}$$

Note that these equations have been written as a stochastic process description, one sample of which will correspond to a single operation of the filter: this is done since we want to evaluate the ensemble average performance of the filter. In terms of this algorithm, the estimated values of the critical quantities of interest at sample time t_i are:

$$\hat{\mathbf{y}}(t_i{}^-) = \mathbf{C}(t_i)\hat{\mathbf{x}}(t_i{}^-) \tag{6-76a}$$

$$\hat{\mathbf{y}}(t_i{}^+) = \mathbf{C}(t_i)\hat{\mathbf{x}}(t_i{}^+) \tag{6-76b}$$

and, if we wanted these estimates any time in the sample period $[t_{i-1}, t_i)$, (6-73) could be used to generate

$$\hat{\mathbf{y}}(t) = \mathbf{C}(t)\hat{\mathbf{x}}(t/t_{i-1}) \tag{6-76c}$$

Since a Kalman filter will often be implemented in an indirect feedback configuration, in which error state estimates are fed back to the actual system to try to correct it, the preceding truth model and filter relations will now be modified to allow performance analysis of such a configuration. One type of feedback discussed previously is the *impulsive control or discrete-time reset*. Quantities maintained as contents of computer memory locations (or outputs of integrators) can be changed instantaneously based upon the error estimate $\hat{\mathbf{x}}(t_i{}^+)$. Examples of variables controlled in this manner in aided inertial systems are position, velocity, and direction cosine matrix elements for attitude information. Maximum rate torquing of the INS platform is commanded to remove estimated misalignments; although this is not an instantaneous change, it is accomplished rapidly enough compared to the filter iteration period that it is well approximated as impulsive. Once $\hat{\mathbf{x}}(t_i{}^+)$ is computed, the reset control is calculated as a function (assumed linear) of it, $\mathbf{D}_t(t_i)\hat{\mathbf{x}}(t_i{}^+)$, where the notation calls out that this is a discrete-time control that affects the true system (truth model). After application of the control, the truth model state process description becomes

$$\mathbf{x}_t(t_i{}^c) = \mathbf{x}_t(t_i) - \mathbf{D}_t(t_i)\hat{\mathbf{x}}(t_i{}^+) \tag{6-77}$$

The filter should be "told" that this feedback to the system has occurred, so its state estimate is modified as

$$\hat{\mathbf{x}}(t_i^{+c}) = \hat{\mathbf{x}}(t_i^{+}) - \mathbf{D}(t_i)\hat{\mathbf{x}}(t_i^{+}) = [\mathbf{I} - \mathbf{D}(t_i)]\hat{\mathbf{x}}(t_i^{+}) \qquad (6\text{-}78)$$

where the n-by-n $\mathbf{D}(t_i)$ is meant to model the effect of feedback through the actual n_t-by-n gains $\mathbf{D}_t(t_i)$ into the system. This $\hat{\mathbf{x}}(t_i^{+c})$ replaces $\hat{\mathbf{x}}(t_i^{+})$ as the initial condition for the next time propagation.

Some true system variables are controlled over the entire sample period rather than impulsively. As discussed in Section 6.5, gyro drift rate compensation is achieved by torquing the gyro with a constant (analog signal or pulse rate) control over a sample period $[t_{i-1}, t_i)$, proportional to the negative of the drift rate estimate $\hat{\varepsilon}(t_{i-1}^{+})$ obtained at the beginning of the interval. This form of feedback could be implemented directly into the performance analysis, but a simpler and good approximate model is to treat the feedback as *continuous control*, $\mathbf{X}_t(t)\hat{\mathbf{x}}(t/t_{i-1})$. Thus, the truth model equation (6-58) is modified to

$$\dot{\mathbf{x}}_t(t) = \mathbf{F}_t(t)\mathbf{x}_t(t) - \mathbf{X}_t(t)\hat{\mathbf{x}}(t/t_{i-1}) + \mathbf{G}_t(t)\mathbf{w}_t(t) \qquad (6\text{-}79)$$

Again, the filter is to be informed of such feedback, so (6-73a) becomes

$$\dot{\hat{\mathbf{x}}}(t/t_{i-1}) = \mathbf{F}(t)\hat{\mathbf{x}}(t/t_{i-1}) - \mathbf{X}(t)\hat{\mathbf{x}}(t/t_{i-1}) = [\mathbf{F}(t) - \mathbf{X}(t)]\hat{\mathbf{x}}(t/t_{i-1}) \quad (6\text{-}80)$$

Similarly, $\mathbf{F}(t)$ is replaced by $[\mathbf{F}(t) - \mathbf{X}(t)]$ in (6-73b), or equivalently, the state transition matrix that appears in (6-72) is associated with $[\mathbf{F}(t) - \mathbf{X}(t)]$ rather than $\mathbf{F}(t)$.

At this point, we have described the truth model and Kalman filter that appear in Fig. 6.18. For convenience, we now define the augmented state vector process $\mathbf{x}_a(\cdot, \cdot)$ for this entire configuration: for any time in the interval $[t_{i-1}, t_i)$,

$$\mathbf{x}_a(\cdot, \cdot) = \left[\begin{array}{c} \mathbf{x}_t(\cdot, \cdot) \\ \hline \hat{\mathbf{x}}(\cdot/t_{i-1}, \cdot) \end{array}\right] \qquad (6\text{-}81)$$

i.e., an $(n_t + n)$-vector stochastic process with partitions of the truth model and filter states, respectively. From (6-79) and (6-80), the augmented state vector *time propagation* relation is:

$$\dot{\mathbf{x}}_a(t) = \mathbf{F}_a(t)\mathbf{x}_a(t) + \mathbf{G}_a(t)\mathbf{w}_t(t) \qquad (6\text{-}82)$$

with

$$\mathbf{F}_a(t) = \left[\begin{array}{c|c} \mathbf{F}_t(t) & -\mathbf{X}_t(t) \\ \hline 0 & [\mathbf{F}(t) - \mathbf{X}(t)] \end{array}\right] \qquad \mathbf{G}_a(t) = \left[\begin{array}{c} \mathbf{G}_t(t) \\ \hline 0 \end{array}\right] \qquad (6\text{-}83)$$

solved forward from time t_{i-1} with the initial conditions

$$\mathbf{x}_a(t_{i-1}^{+c}) = \left[\begin{array}{c} \mathbf{x}_t(t_{i-1}^{c}) \\ \hline \hat{\mathbf{x}}(t_{i-1}^{+c}) \end{array}\right] \qquad (6\text{-}84)$$

This can be written equivalently through the discrete-time solution model:

$$\mathbf{x}_a(t_i^-) = \boldsymbol{\Phi}_a(t_i, t_{i-1})\mathbf{x}_a(t_{i-1}^{+c}) + \mathbf{w}_{da}(t_{i-1}) \tag{6-85}$$

where $\boldsymbol{\Phi}_a(t_i, t_{i-1})$ is the $(n_t + n)$-by-$(n_t + n)$ matrix that satisfies

$$\dot{\boldsymbol{\Phi}}_a(t, t_{i-1}) = \mathbf{F}_a(t)\boldsymbol{\Phi}_a(t, t_{i-1}) \tag{6-86a}$$

$$\boldsymbol{\Phi}_a(t_{i-1}, t_{i-1}) = \mathbf{I} \tag{6-86b}$$

and $\mathbf{w}_{da}(\cdot, \cdot)$ is a discrete-time $(n_t + n)$-vector white Gaussian noise process with mean zero and covariance:

$$E\{\mathbf{w}_{da}(t_{i-1})\mathbf{w}_{da}^{\mathrm{T}}(t_{i-1})\} = \mathbf{Q}_{da}(t_{i-1})$$
$$= \int_{t_{i-1}}^{t_i} \boldsymbol{\Phi}_a(t_i, \tau)\mathbf{G}_a(\tau)\mathbf{Q}_1(\tau)\mathbf{G}_a^{\mathrm{T}}(\tau)\boldsymbol{\Phi}_a^{\mathrm{T}}(t_i, \tau)\, d\tau \tag{6-87}$$

To generate the *measurement update* relations, first of all the truth model state is unchanged:

$$\mathbf{x}_t(t_i^+) = \mathbf{x}_t(t_i^-) \tag{6-88}$$

The filter update can be written as:

$$\begin{aligned}
\hat{\mathbf{x}}(t_i^+) &= \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[\mathbf{z}_t(t_i) - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \\
&= [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)\mathbf{z}_t(t_i) \\
&= [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)\mathbf{H}_t(t_i)\mathbf{x}_t(t_i) + \mathbf{K}(t_i)\mathbf{v}_t(t_i) \tag{6-89}
\end{aligned}$$

Equations (6-88) and (6-89) yield an augmented state vector measurement update as

$$\mathbf{x}_a(t_i^+) = \mathbf{A}_a(t_i)\mathbf{x}_a(t_i^-) + \mathbf{K}_a(t_i)\mathbf{v}_t(t_i) \tag{6-90}$$

where

$$\mathbf{A}_a(t_i) = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{K}(t_i)\mathbf{H}_t(t_i) & [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)] \end{array}\right] \tag{6-91a}$$

$$\mathbf{K}_a(t_i) = \left[\begin{array}{c} \mathbf{0} \\ \hline \mathbf{K}(t_i) \end{array}\right] \tag{6-91b}$$

The *impulsive (reset) control update* can be described, using (6-77) and (6-78), as

$$\mathbf{x}_a(t_i^{+c}) = \mathbf{D}_a(t_i)\mathbf{x}_a(t_i^+) \tag{6-92}$$

where

$$\mathbf{D}_a(t_i) = \left[\begin{array}{c|c} \mathbf{I} & -\mathbf{D}_t(t_i) \\ \hline \mathbf{0} & [\mathbf{I} - \mathbf{D}(t_i)] \end{array}\right] \tag{6-93}$$

Note that if feedback is not employed, $\mathbf{D}_a(t_i)$ is simply an $(n_t + n)$-by-$(n_t + n)$ identity matrix. Initial conditions for these relations are:

$$\mathbf{x}_a(t_0) = \left[\frac{\mathbf{x}_t(t_0)}{\hat{\mathbf{x}}(t_0)} \right] = \left[\frac{\mathbf{x}_t(t_0)}{\mathbf{0}} \right] \tag{6-94}$$

Finally, the output $\mathbf{e}_t(\cdot, \cdot)$ in Fig. 6.16, the error committed by the filter in estimating the essential quantities, can be generated from the augmented state vector at any time $t \in T$ as

$$\mathbf{e}_t(t) = \mathbf{C}_a(t)\mathbf{x}_a(t) \tag{6-95}$$

where

$$\mathbf{C}_a(t) = \left[-\mathbf{C}_t(t) \mid \mathbf{C}(t) \right] \tag{6-96}$$

In a *Monte Carlo study*, these relationships are used to generate individual *samples* of stochastic processes, employing random number generators to obtain each simulation (see Problem 7.14 for a discussion of process sample generation). A set of initial conditions is established through a realization of (6-94), time propagation conducted through a sample from (6-82) or (6-85), and measurement and control updates via (6-90) and (6-92). Thus, a sample of the estimation error process is computed from (6-95). The output from a single computer run of a Monte Carlo study is a single sample $\mathbf{e}_t(t, \omega_1)$ for all time t of interest. A second run produces $\mathbf{e}_t(t, \omega_2)$ for all t, and so forth, as portrayed in Fig. 6.19. A statistical description of this error process is achieved by computing sample statistics, averaging over the number of runs conducted.

The equations for a *covariance analysis* [8, 17, 18, 31] are readily obtained from these relations. Since all processes are assumed to be zero mean, the covariances of the augmented state and estimation error can be defined as

$$\mathbf{P}_a(t) = E\{\mathbf{x}_a(t)\mathbf{x}_a^T(t)\} \tag{6-97}$$

$$\mathbf{P}_e(t) = E\{\mathbf{e}_t(t)\mathbf{e}_t^T(t)\} \tag{6-98}$$

The time history of $\mathbf{P}_e(t)$ is the desired output and $\mathbf{P}_a(t)$ is calculated as a means of obtaining this result. The appropriate *initial conditions* are obtained from (6-94) as

$$\mathbf{P}_a(t_0) = E\{\mathbf{x}_a(t_0)\mathbf{x}_a^T(t_0)\} = \left[\begin{array}{c|c} \mathbf{P}_{t0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{P}_0 \end{array} \right] \tag{6-99}$$

Propagating between sample times is accomplished by integrating

$$\dot{\mathbf{P}}_a(t) = \mathbf{F}_a(t)\mathbf{P}_a(t) + \mathbf{P}_a(t)\mathbf{F}_a^T(t) + \mathbf{G}_a(t)\mathbf{Q}_t(t)\mathbf{G}_a^T(t) \tag{6-100}$$

forward from time t_{i-1}, with initial conditions as $\mathbf{P}_a(t_{i-1}^{+c})$, to time t_i, as seen from (6-82)–(6-84). This can be expressed equivalently, from (6-85)–(6-87), as

$$\mathbf{P}_a(t_i^-) = \mathbf{\Phi}_a(t_i, t_{i-1})\mathbf{P}_a(t_{i-1}^{+c})\mathbf{\Phi}_a^T(t_i, t_{i-1}) + \mathbf{Q}_{da}(t_{i-1}) \tag{6-101}$$
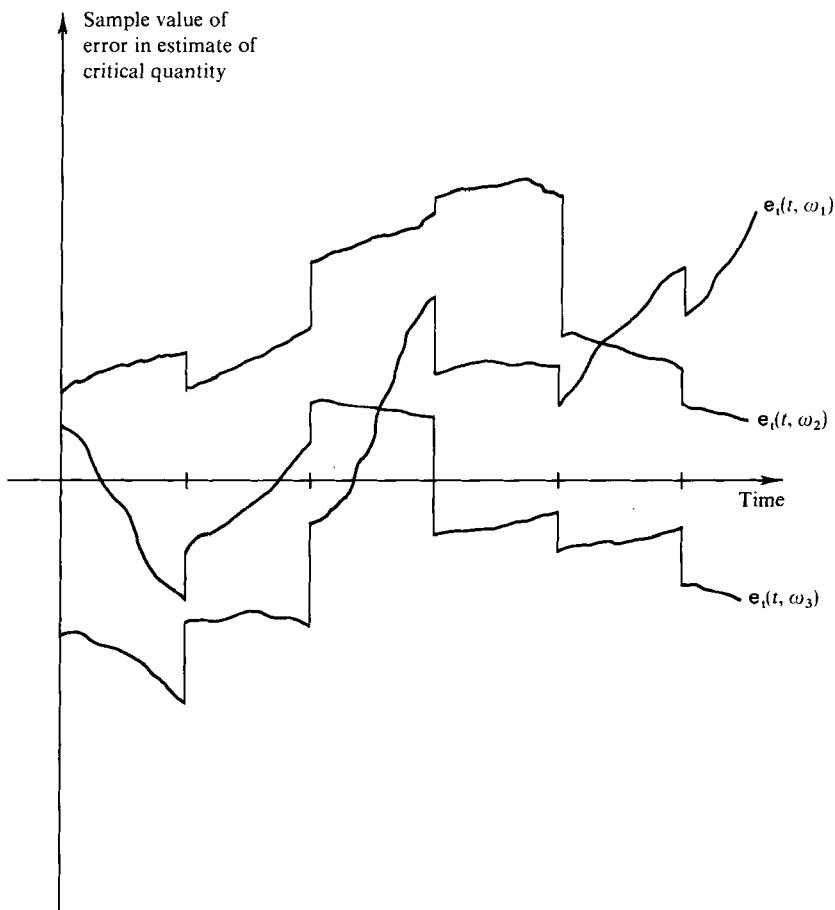
FIG. 6.19   Outputs of three Monte Carlo simulation runs.

The *measurement update* relation derived from (6-90) and (6-91) is

$$\mathbf{P}_a(t_i^+) = \mathbf{A}_a(t_i)\mathbf{P}_a(t_i^-)\mathbf{A}_a^T(t_i) + \mathbf{K}_a(t_i)\mathbf{R}_t(t_i)\mathbf{K}_a^T(t_i) \tag{6-102}$$

From (6-92) and (6-93), the *impulsive (reset) control update* is

$$\mathbf{P}_a(t_i^{+c}) = \mathbf{D}_a(t_i)\mathbf{P}_a(t_i^+)\mathbf{D}_a^T(t_i) \tag{6-103}$$

As the state covariance recursion is generated, the desired *error covariance* can be obtained at any time of interest, using (6-95) and (6-96), as

$$\mathbf{P}_e(t) = \mathbf{C}_a(t)\mathbf{P}_a(t)\mathbf{C}_a^T(t) \tag{6-104}$$

Because the augmented system is a linear system driven by white Gaussian noise, the covariance relations are independent of the actual measurement

time history, so it is possible to perform this covariance analysis *a priori, without* resorting to noise generator simulation of process samples.

To conduct a covariance analysis, one must explicitly define the structure and uncertainties of the truth model (F_t or Φ_t, G_t, H_t, Q_t, and R_t time histories and P_{t0}) and of the design model upon which the filter is based (F or Φ, G, H, Q, and R time histories and P_0). Note that the only effect of $Q(\cdot)$ and $R(\cdot)$ in these equations is to establish the filter gain $K(\cdot)$: the design model $w(\cdot,\cdot)$ and $v(\cdot,\cdot)$ do not directly affect the spreading of samples of the $\hat{x}(\cdot,\cdot)$ process. Moreover, a filter modification which incorporates approximated gains instead can be analyzed readily.

One particular use of a covariance performance (sensitivity) analysis is a systematic approach to the *filter tuning* [13, 15, 16, 20, 24, 27, 33, 37–39, 50, 51] process. The basic objective of filter tuning is to achieve the best possible estimation performance from a filter of specified structural form, i.e., totally specified except for P_0 and the time histories of Q and R. These covariances not only account for actual noises and disturbances in the physical system, but also are a means of declaring how adequately the assumed model represents the "real world" system. The simpler and less accurate the model, the stronger the noise strengths should be set, but it is difficult if not impossible to specify best numerical values a priori.

In the tuning process, the filter structure and the entire truth model remain fixed. For one set of values of P_0 and time histories of Q and R in the filter, the covariance analysis provides the time history of the "true" rms errors in estimates of critical quantities committed by the filter, obtained from the square roots of diagonal terms of P_e. Another set of these noise parameters can be chosen, thereby generating another time history of rms estimation errors. This procedure can be repeated until the "best" choice of parameters is found, best in the sense that filter errors are minimized. Thus, the tuning process can be considered a numerical optimization problem, a constant parameter optimization if stationary statistics are assumed, or a significantly more difficult function optimization if stationarity is not assumed. As such, it can be solved by automatic search methods or by manual calculation: using physical insights to propose changes in the noise covariances, and continuing to vary them until the performance no longer improves. Manual "optimization" is more prevalent in practice. Basically, the P_0 matrix is the determining factor in the initial transient performance of the filter, whereas the Q and R histories dictate the longer term ("steady state" if time-invariant system and stationary noise models apply) performance and time duration of transients.

When performing the filter tuning, it is useful to compare the actual rms errors committed by the filter to the filter's own representation of its errors: the P covariance time history computed as an integral part of the estimator algorithm. This time history is also directly available as an output from the covariance analysis [$P_e(t)$ would be directly comparable to $C(t)P(t)C^T(t)$, but

all filter channels are usually observed individually for tuning purposes, so here we consider $\mathbf{C}(t) \triangleq \mathbf{I}$ and $\mathbf{C}_a(t)$ defined appropriately]. One would want the filter to perform as well as it "believes" it is performing. If, due to mistuned noise parameters, the filter underestimates its own errors, it will not "look hard enough" at the measurements, with resulting filter divergence problems if the discrepancy is significant enough [12, 43, 44], as depicted in Fig. 6.20a.
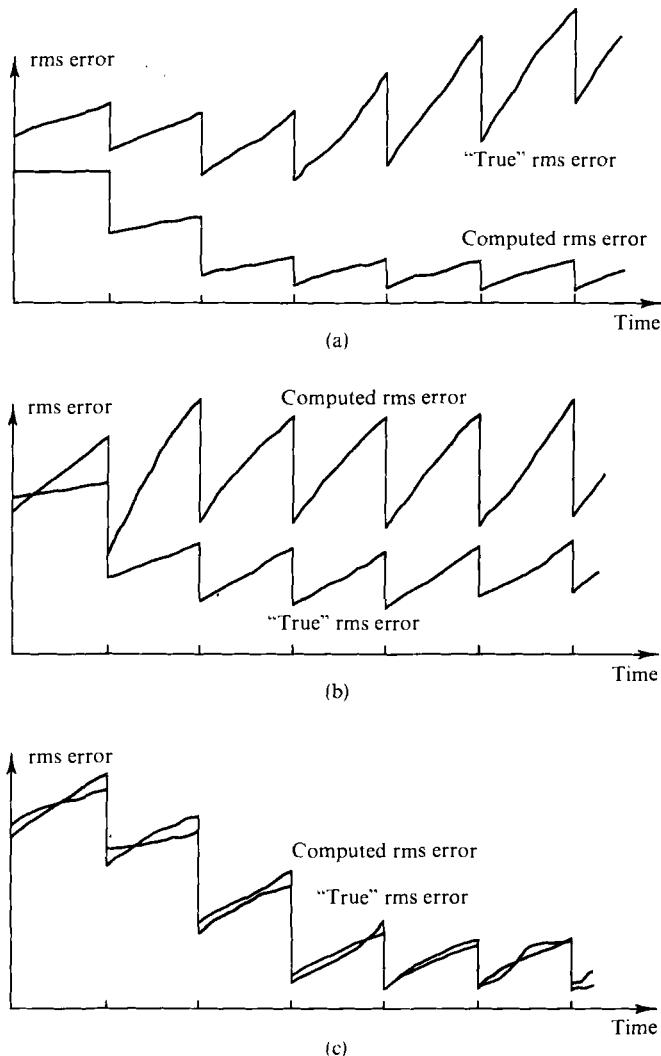
FIG. 6.20   Filter tuning through covariance analysis. (a) Filter underestimates its own errors: divergence. (b) Filter overestimates its own errors: tracking of measurement noise. (c) Well-tuned filter.

On the other hand, if the filter's internally generated covariance overestimates the errors being committed, then the filter weights the measurements too heavily, expending too much effort to track the noisy data and not exploiting the benefits of its internal model enough, as seen in Fig. 6.20b. By choosing the noise parameters so that the overall time histories of actual rms errors and the square roots of the filter **P** matrix diagonal terms match well, the actual estimation errors are effectively reduced at the same time. This is portrayed in Fig. 6.20c: note that the true rms error is lower than in the two preceding plots. Allowing the filter to overestimate its own errors slightly, so as to minimize the likelihood of divergence, is a commonly adopted means of generating a conservative filter design.

Other means of filter tuning are possible as well. If truth model noise (and structure) parameters are known only with some uncertainty, it is possible to tune a reduced order filter by a game theoretic (minimax) approach, in which the uncertain parameters "try" to maximize the estimation errors and the filter design parameters "try" to minimize it. The result is a single filter design with acceptable performance over the entire range of uncertain parameter values [9], without increased online computation required, as would be the case for an adaptive estimator.

Once a particular filter has been tuned, an *error budget* [1, 15, 16, 20, 24, 27, 33, 37–39, 50, 51] can be established. Essentially, this consists of repeated covariance analyses in which the error sources (or small groups of sources) in the truth model are "turned on" individually to determine the separate effects of these sources. At particular times of interest, the rms errors in estimates of quantities of interest due only to a single source of error are recorded. For instance, Fig. 6.21 plots a north position rms error budget for a particular filter considered for a Doppler-aided INS application, at a point following a long period of overwater flight (a worst-case scenario for such a navigation system). This information is useful for a number of purposes. If the filter under test were based upon the full-scale truth model, this would indicate that it is most essential for the filter to model INS gyro drift rate (three single-state shaping filters) and wideband Doppler noise (modeled as white noise): it suggests a filter that neglects the states modeling the other contributions and increasing the strength of noises to account for such deletion. If the filter under test were such a practical design, and tuned properly, then this error budget would indicate navigation errors being dominated by gyro drift rate and Doppler noise: if hardware were to be improved, the most cost-effective improvement would be an INS with better gyro drift characteristics and/or a better precision Doppler. Note that, because of the linearity of the covariance analysis equations, the total navigation rms error can be found as the root-sum-square of the individual independent contributions.

Besides tuning or generating an error budget for a single filter, a covariance analysis provides an effective means of conducting a *tradeoff analysis* among

Source:

INS Gyro
drift rate

Doppler
noise

Other
gyro errors

Gravity
anomalies

Doppler
bias

Accelerometer
noise

Accelerometer
bias

Altimeter
noise

Contribution to
rms north
position error (ft)
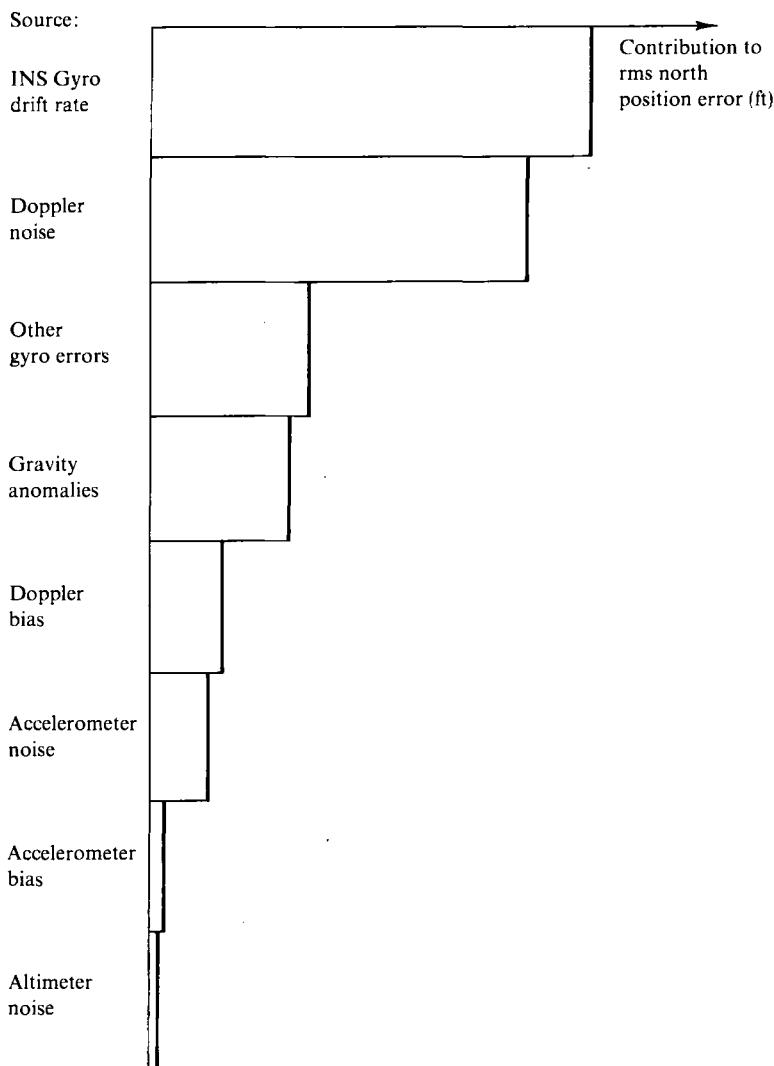
FIG. 6.21   Navigation sensor error budget subsequent to long overwater flight.

the various proposed designs. By explicitly evaluating the performance capabilities and computer burden of each design, the engineer has obtained the information necessary to make a rational selection of the filter most appropriate to his particular application. *Sensitivity to parameter variations* in the filter or system hardware (i.e., the truth model) can be obtained in a straightforward manner by repeated covariance analyses, or adjoint methods [7] can be em-

ployed to describe local sensitivity to small variations in many parameters simultaneously.

Although a covariance performance analysis is computationally more efficient than a Monte Carlo study and so should be exploited, there is a definite advantage to employing the Monte Carlo approach in addition. First of all, a Monte Carlo study encompasses a *system simulation* in which the *actual filter algorithm* is embedded. As such, portions of the simulation can be *replaced by actual data or hardware* as it becomes available. Moreover, *sign errors* in the filter algorithm that may not be readily apparent in a covariance analysis due to squaring effects become evident in the Monte Carlo study: if the two performance analyses are based on the same models and disagree significantly in predicted estimation accuracy, sign errors are to be suspected. Finally, effects of *nonlinearities* such as device saturation or neglected terms in attaining linear perturbation equations cannot be evaluated by a covariance approach, and must be investigated through Monte Carlo means.

6.9   SYSTEMATIC DESIGN PROCEDURE

As can be surmised from the preceding sections, Kalman filter development requires an iterative proposing of filter modifications and evaluating of each version's capabilities of meeting performance requirements (rms estimation errors, etc.) and practical constraints (computation time, storage, sequencing, cost, etc.). A *systematic design procedure* would be conducted in the following manner:

(1)   Develop a "truth model" (a complete, complex mathematical model that portrays true system behavior very accurately) and validate with laboratory and operational test data (this serves as the basis for evaluating all prospective designs, so one must establish its validity with the utmost confidence); if nonlinear, linearize it for later covariance analyses.

(2)   Generate the Kalman filter based upon the "truth" model as a "benchmark" of performance and analyze its capabilities. (If the truth model is linear, the filter-generated covariance provides a valid performance analysis; if nonlinear, Monte Carlo evaluations are necessary.)

(3)   Propose simplified, reduced order system models by removing and combining states associated with nondominant effects, deleting weak cross-coupling terms, employing approximations, etc. (this requires substantial physical insights into the problem at hand), and develop the simplified filters based on each model; also consider approximated filter structures.

(4)   Conduct a covariance performance analysis of each proposed Kalman filter being driven by measurements derived from the truth model of the real system; as an iteration within this step, "tune" each filter to provide best possible performance from each.

(5)   Generate a Monte Carlo analysis of designs from the preceding step that show most promise.

(6)   Conduct a performance/computer loading tradeoff analysis and select a design; investigate square root and other forms of implementing the chosen design.

(7)   Implement the chosen design on the online computer to be used in the final system.

(8)   Perform checkout, final tuning, and operational test of the filter.

Through such a procedure, a logical decision process based on sufficient empirical data is incorporated into the design. As a result, the final implementation should perform as well as predicted in earlier stages of design. (If the implementation is error free and performance is not as predicted, this indicates that the original declaration of the truth model was faulty.) Although the procedure consumes much time and effort, the overall design risk is less than that associated with making design decisions with less supportive analysis, in which inadequate performance is not detected until operational test. It is, in fact, a cost-effective procedure.

6.10   INS AIDED BY NAVIGATION SATELLITES

This section illustrates use of the design procedure of Section 6.9 to generate a practical Kalman filter for a full-scale example of aiding an inertial system with position data provided by navigation satellites [5, 10]. First the truth model will be delineated and the performance of the filter based on the truth model portrayed. Based on insights gained from the physics of the problem and this covariance analysis, reduced order filters will be suggested and analyzed. One parameter of distinct interest for this problem is the filter sample period, and sensitivity to its variation will be demonstrated.

The Global Positioning System (GPS) currently under development consists of 24 navigation satellites, placed eight in each of three different orbits. Each satellite contains a transmitter, receiver, and a quartz crystal oscillator "clock." Periodically a ground tracking network measures and updates the ephemeris and satellite clock phase and frequency in order to maintain synchronization of all satellite clocks. The satellites in turn continually transmit this information together with a range code, formatted such that each satellite signal can be distinguished from the others by the user. By means of a correlation detector, the time (phase) shift between each satellite signal and the user's unsynchronized clock is measured in his receiver, to provide an indication of range from the satellite to the user. If the user has an INS, its position indication and the satellite emphemeris data can be used to compute an INS-indicated range to the satellite. The difference of these two range indications, called "range

divergence," serves as an input to an indirect feedback Kalman filter to yield an integrated GPS-aided inertial navigation system. Four such range divergence signals become available at each filter sample time, i.e., from four separate satellites, in order to correct the three components of position and clock phase difference (a dominant effect which enters the model directly as a position error).

A scalar range divergence measurement associated with satellite j ($j = 1, 2, 3, 4$) can be written in terms of INS-indicated range $R_{j\text{-INS}}$, satellite-indicated range $R_{j\text{-sat}}$, true range $R_{j\text{-true}}$, and associated errors $\delta R_{j\text{-INS}}$ and $\delta R_{j\text{-sat}}$ as:

$$
\begin{aligned}
z_j(t_i) &= R_{j\text{-INS}}(t_i) - R_{j\text{-sat}}(t_i) \\
&= \left[R_{j\text{-true}}(t_i) + \delta R_{j\text{-INS}}(t_i) \right] - \left[R_{j\text{-true}}(t_i) + \delta R_{j\text{-sat}}(t_i) \right] \\
&= \delta R_{j\text{-INS}}(t_i) - \delta R_{j\text{-sat}}(t_i)
\end{aligned} \tag{6-105}
$$

Since INS-indicated range is based upon INS-indicated position of the user vehicle and satellite-provided ephemeris data on its own current position, $\delta R_{j\text{-INS}}(t_i)$ is due to errors in both these sources. However, the satellite orbital parameters are very precisely updated by the ground tracking network and are negligible (any small uncertainties due to this effect can be accounted for by increasing the satellite clock phase error). Let the INS platform instrument an east–north–up coordinate frame so that the INS position errors are $\delta \mathbf{r} = [\delta r_e, \delta r_n, \delta r_u]^T$. Then if $\mathbf{i}_j = [i_{je}, i_{jn}, i_{ju}]^T$ is the unit vector direction from the user to satellite j, expressed in east–north–up coordinates, the error $\delta R_{j\text{-INS}}(t_i)$ can be written as

$$
\begin{aligned}
\delta R_{j\text{-INS}}(t_i) &= -\mathbf{i}_j(t_i)^T \, \delta \mathbf{r}(t_i) \\
&= \left[-i_{je}(t_i) \right] \delta r_e(t_i) + \left[-i_{jn}(t_i) \right] \delta r_n(t_i) + \left[-i_{ju}(t_i) \right] \delta r_u(t_i) \tag{6-106}
\end{aligned}
$$

The range measurement error model, validated through empirical data, is

$$
\delta R_{j\text{-sat}}(t_i) = c \, \delta T_u(t_i) - c \, \delta T_{sj}(t_i) + b_{rj}(t_i) - v_j(t_i) \tag{6-107}
$$

where c is the speed of light, $\delta T_u(t_i)$ is user clock phase error, $\delta T_{sj}(t_i)$ is satellite j clock phase error, including error due to ionospheric delay, $b_{rj}(t_i)$ is a small range bias term accounting for tropospheric delay and uncertainty in the speed of light, and $v_j(t_i)$ is white Gaussian measurement noise modeling high frequency corruption of satellite signal and quantization error. Substitution of (6-106) and (6-107) into (6-105) yields the $z_j(t_i)$ measurement relation for the truth model.

The truth model dynamics model is based upon the Pinson error model [40, 41, 52] for the three position errors ($\delta r_e, \delta r_n, \delta r_u$), three velocity errors ($\delta v_e, \delta v_n, \delta v_u$), and three attitude errors (ψ_e, ψ_n, ψ_u) in an INS, specified by the 9-by-9 $\mathbf{F}(\cdot)$ given in Fig. 6.22 and driven by sources of uncertainty.

The time rates of change of the velocity errors are driven by accelerometer errors as given in Example 6.1 of Section 6.7: three accelerometers with six

$$
\begin{array}{c}
\dot{\delta r_e}: \\
\dot{\delta r_n}: \\
\dot{\delta r_u}: \\
\dot{\delta v_e}: \\
\dot{\delta v_n}: \\
\dot{\delta v_u}: \\
\dot\psi_e: \\
\dot\psi_n: \\
\dot\psi_u:
\end{array}
$$

	δr_e	δr_n	δr_u	δv_e	δv_n	δv_u	ψ_e	ψ_n	ψ_u
$\dot{\delta r_e}:$	0	0	0	1	0	0	0	0	0
$\dot{\delta r_n}:$	0	0	0	0	1	0	0	0	0
$\dot{\delta r_u}:$	0	0	0	0	0	1	0	0	0
$\dot{\delta v_e}:$	$(A_z/R) - \omega_s^2 + \omega^2 - \Omega^2 - (\omega_x + \Omega_x)\rho_x$	$\dot\rho_z - (\omega_y + \Omega_y)\rho_x$	$-\dot\rho_y - (\omega_z + \Omega_z)\rho_x$	0	$2\omega_z$	$-2\omega_y$	0	$-A_z$	A_y
$\dot{\delta v_n}:$	$-\dot\rho_z - (\omega_x + \Omega_x)\rho_y$	$(A_z/R) - \omega_s^2 + \omega^2 - \Omega^2 - (\omega_y + \Omega_y)\rho_y$	$\dot\rho_x - (\omega_z + \Omega_z)\rho_y$	$-2\omega_z$	0	$2\omega_x$	A_z	0	$-A_x$
$\dot{\delta v_u}:$	$\dot\rho_y - (A_x/R) - (\omega_x + \Omega_x)\rho_z$	$-\dot\rho_x - (A_y/R) - (\omega_y + \Omega_y)\rho_z$	$\omega^2 - \Omega^2 + 2\omega_s^2 - (\omega_z + \Omega_z)\rho_z$	$2\omega_y$	$-2\omega_x$	0	$-A_y$	A_x	0
$\dot\psi_e:$	$-\omega_z/R$	0	0	0	$1/R$	0	0	ω_z	$-\omega_y$
$\dot\psi_n:$	0	$-\omega_z/R$	0	$1/R$	0	0	$-\omega_z$	0	ω_x
$\dot\psi_u:$	ω_x/R	ω_y/R	0	0	0	0	ω_y	$-\omega_x$	0

FIG. 6.22   9-by-9 $\mathbf{F}$ matrix of Pinson INS error model. R = earth radius, $\mathbf{A}$ = acceleration, ρ = angular rate from earth-fixed coordinates to platform coordinates, Ω = earth angular rate relative to inertial coordinates, ω = angular rate from inertial coordinates to platform coordinates ($\omega = \Omega + \rho$), $\omega^2 = \omega^T\omega$, $\omega_s^2 = g/R$ = Schuler frequency squared; subscripts x, y, and z correspond to vector components in the east, north, and up directions, respectively.

error states per sensor, yielding 18 augmented states. Also driving these velocity errors are gravity anomalies: three exponentially distance-correlated states to model the difference between the geoid and the ellipse assumed in the navigation equations; these enter because computed gravity is subtracted from the specific force outputs of the accelerometers to yield vehicle acceleration. Exponentially distance-correlated processes are the outputs of first order lags driven by white noise, with the lag time constant T equal to $[d/V]$, where d is the correlation distance and V is the vehicle velocity magnitude. The time rates of change of attitude errors are driven by gyro errors as given in Example 6.2 of Section 6.7: three single-degree-of-freedom gyros with 14 error states per sensor, or 42 additional states. Thus, the INS error model incorporates 72 states.

The mathematical models for both the user clock and the four satellite clocks [5, 10, 21] are structurally identical, but the user clock initial errors are orders of magnitude greater than those of the satellite clocks because of the greater accuracy of the periodically updated satellite clocks. The generic model is portrayed in Fig. 6.23 and is specified by

$$\delta T(t) = c_0 + c_1 t + c_2 t^2 + e_T(t) \tag{6-108}$$
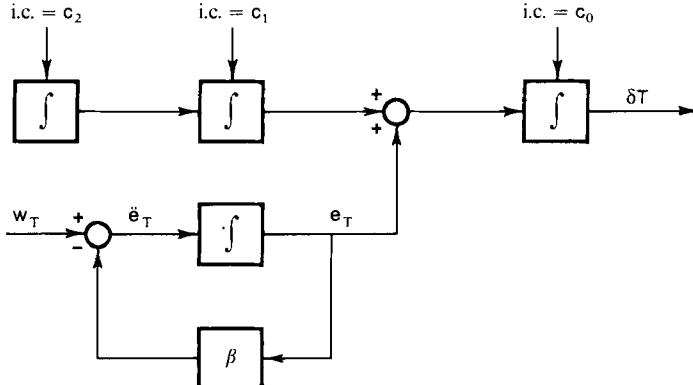
FIG. 6.23   Clock error model.

where c_0, c_1, and c_2 are random variables and $\dot{e}_T(\cdot, \cdot)$ is an exponentially time-correlated noise satisfying

$$\ddot{e}_T(t) = -\beta \dot{e}_T(t) + w_T(t) \tag{6-109}$$

where $w_T(\cdot, \cdot)$ is white Gaussian noise. Thus, each of the five clocks contributes four more states to the system model, for a total of 20 augmented state variables.

Finally, the range biases b_{rj} in (6-107) are modeled as the outputs of undriven integrators. This adds four states, to yield a total truth model state vector of dimension 96.

Truth model parameter values were set so as to represent a state-of-the-art INS being updated every 30 sec by four GPS satellite range measurements. The user vehicle was simulated as flying for one hour at constant speed and altitude over a great circle path. First the Kalman filter based upon the truth model was studied through a covariance performance analysis, concentrating on the nine error states of position, velocity, and attitude indications. Figure
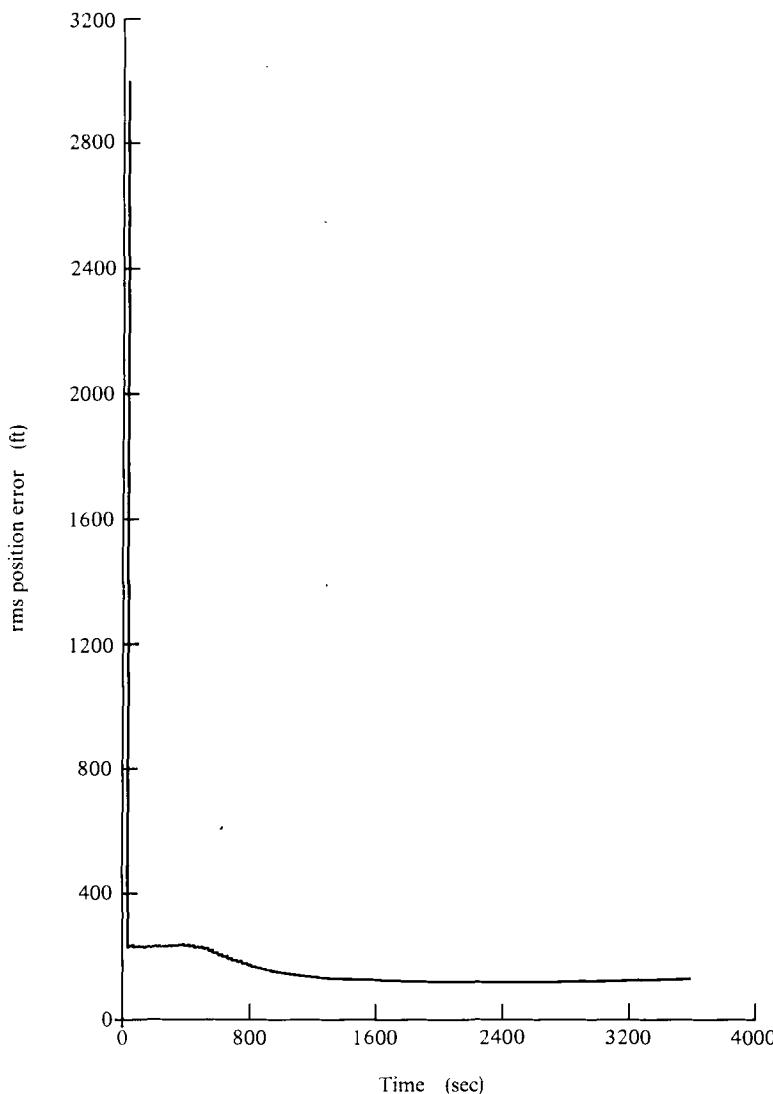
FIG. 6.24a   North rms position error of filter based on truth model.

6.24 presents the north rms position and velocity errors achieved. Note that the position error in Fig. 6.24a starts to grow slightly at the end of one hour's flight, indicative of possible divergence problems for longer flights. This problem is the result of using the same four satellites for all updates: in practice an optimal set of four (ten will be in view at a given time on the average) would be chosen, and analyses have shown this procedure to preclude divergence.
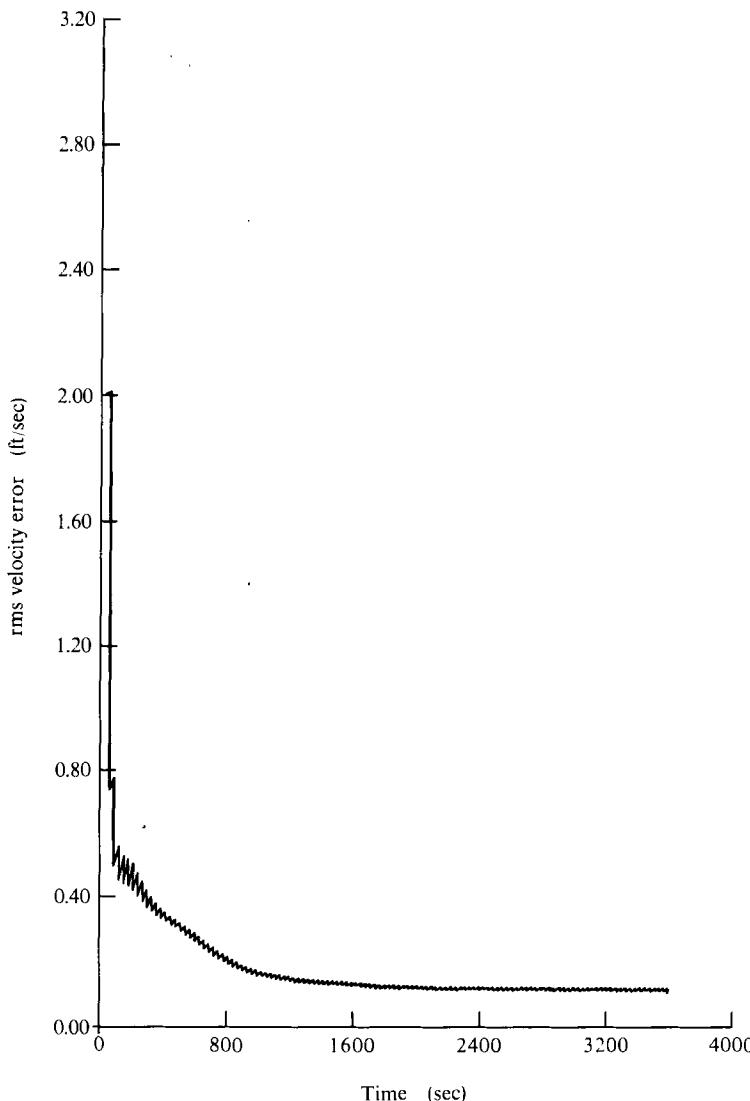
FIG. 6.24b   North rms velocity error of filter based on truth model.

Although a 96-state Kalman filter would be totally impractical, this does serve as a baseline against which to compare other reduced-order simplified designs.

Through error budgets and physical insights, the least significant state variables were deleted or combined, shaping filters reduced in complexity (in the cases of INS gyro and accelerometer models, the gravity anomaly models, and clock $\dot{e}_T$ models, to the limiting case of white noise, requiring no states at all), and remaining noise strengths increased to account for the simplifications. Eventually a 15-state design was attained, being composed of the nine basic INS states $(\delta r_e, \delta r_n, \delta r_u, \delta v_e, \delta v_n, \delta v_u, \psi_e, \psi_n, \psi_u)$, the c_0 clock phase/range error state in (6-108) for each of the four satellites, and both the c_0 and c_1 frequency offset state for the user clock. Moreover, the weak coupling terms in the $\mathbf{F}$ matrix were removed; Schuler frequency terms ($\omega_s{}^2$ and $2\omega_s{}^2$) were retained so that the model applicability would not be restricted to short periods of time, on the order of 30 min. Thus, the upper 9-by-9 partition of the filter $\mathbf{F}$ matrix replaces that given in Fig. 6.22 with the matrix depicted in Fig. 6.25.

$$
\begin{array}{c}
\begin{array}{ccccccccc}
\delta r_e & \delta r_n & \delta r_u & \delta v_e & \delta v_n & \delta v_u & \psi_e & \psi_n & \psi_u
\end{array} \\
\begin{array}{c}
\dot{\delta r}_e: \\
\dot{\delta r}_n: \\
\dot{\delta r}_u: \\
\dot{\delta v}_e: \\
\dot{\delta v}_n: \\
\dot{\delta v}_u: \\
\dot{\psi}_e: \\
\dot{\psi}_n: \\
\dot{\psi}_u:
\end{array}
\begin{bmatrix}
 & & & 1 & & & & & \\
 & & & & 1 & & & & \\
 & & & & & 1 & & & \\
-\omega_s{}^2 & & & & & & & -A_z & A_y \\
 & -\omega_s{}^2 & & & & & A_z & & -A_x \\
 & & 2\omega_s{}^2 & & & & -A_y & A_x & \\
 & & & & & & & \omega_z & -\omega_y \\
 & & & & & & -\omega_z & & \omega_x \\
 & & & & & & \omega_y & -\omega_x &
\end{bmatrix}
\end{array}
$$

FIG. 6.25   Modified 9-by-9 $\mathbf{F}$ matrix of Pinson error model (nonzero elements only).

The performance of this 15-state filter, after being properly tuned, is displayed in Fig. 6.26. Despite the vast reduction in size and complexity of the filter, performance degradation from that of Fig. 6.24 is totally insignificant in position estimation accuracy, while being more apparent in the velocity estimation. However, this design is performing well within specification and provides a good tradeoff of complexity and performance. Range-rate measurements could be used additionally to improve this filter's ability to maintain accurate velocity estimates, if desired.

Some insights suggested a further simplification might be achieved by discarding the four satellite clock phase/range error states and the user clock
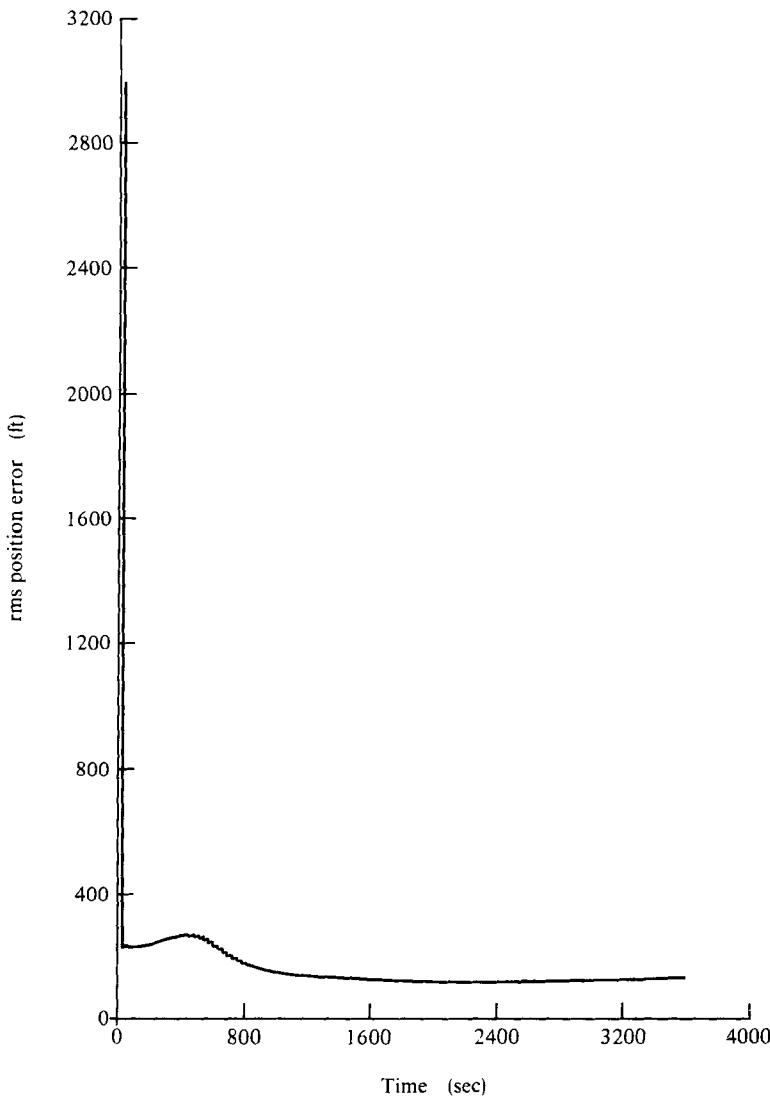
FIG. 6.26a   North rms position error of 15-state simplified filter.

frequency offset error, yielding a ten-state filter. However, performance analysis of this design (after retuning) revealed an unacceptable degradation in estimation capability, as depicted in Fig. 6.27.

Returning attention to the acceptable 15-state design, it was desired to observe the sensitivity of performance to slower update rates. Figure 6.28
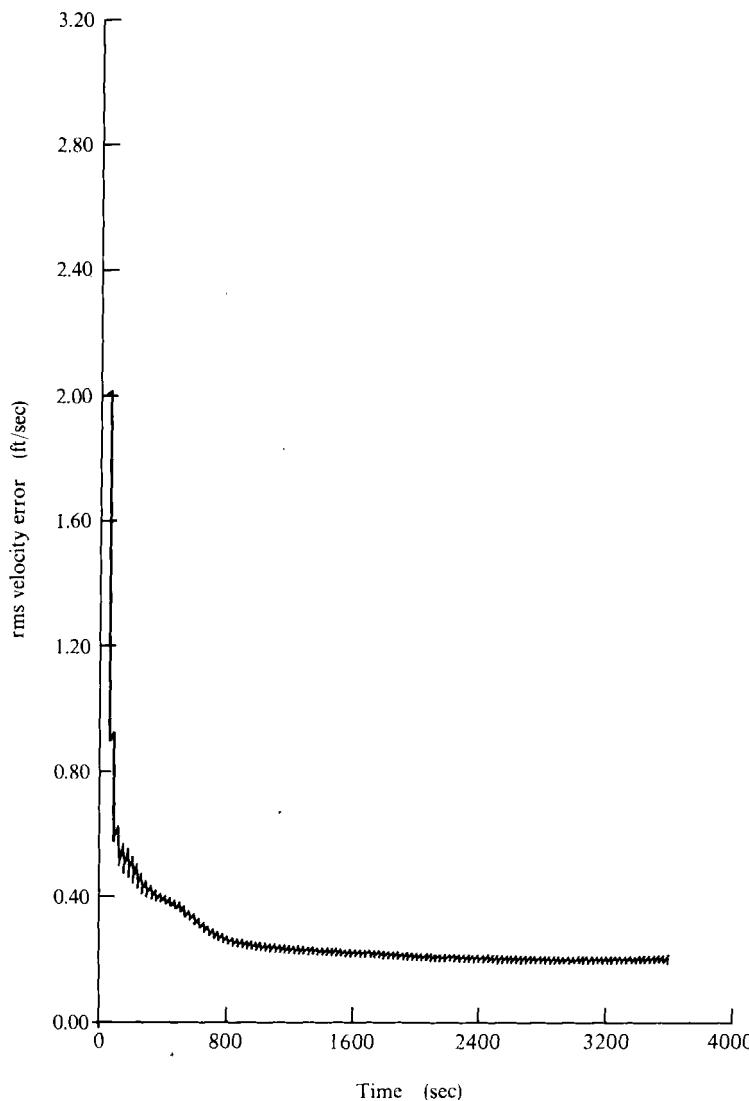
FIG. 6.26b   North rms velocity error of 15-state simplified filter.

indicates that tripling the sample period from 30 to 90 sec still yields acceptable estimation precision.

Thus, covariance performance analysis has been shown to be a versatile tool for initial stages of Kalman filter design and tradeoff studies. Nevertheless, this is but a part of a total systematic design and implementation of an operational filter algorithm.
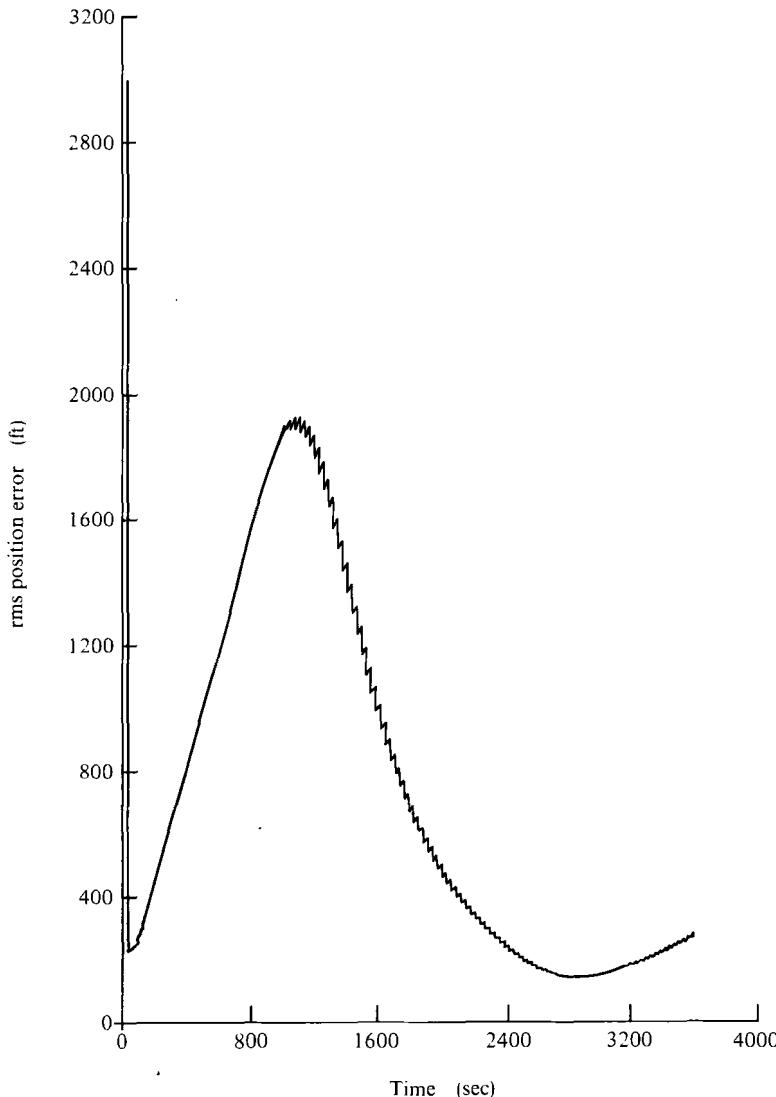
FIG. 6.27a   North rms position error of 10-state simplified filter.

6.11   PRACTICAL ASPECTS OF IMPLEMENTATION[1]

Throughout the development of a filter algorithm, one must be aware of the constraints imposed by the computer in which the software will reside:

[1] See References [28, 29, 31, 35, 46–48].
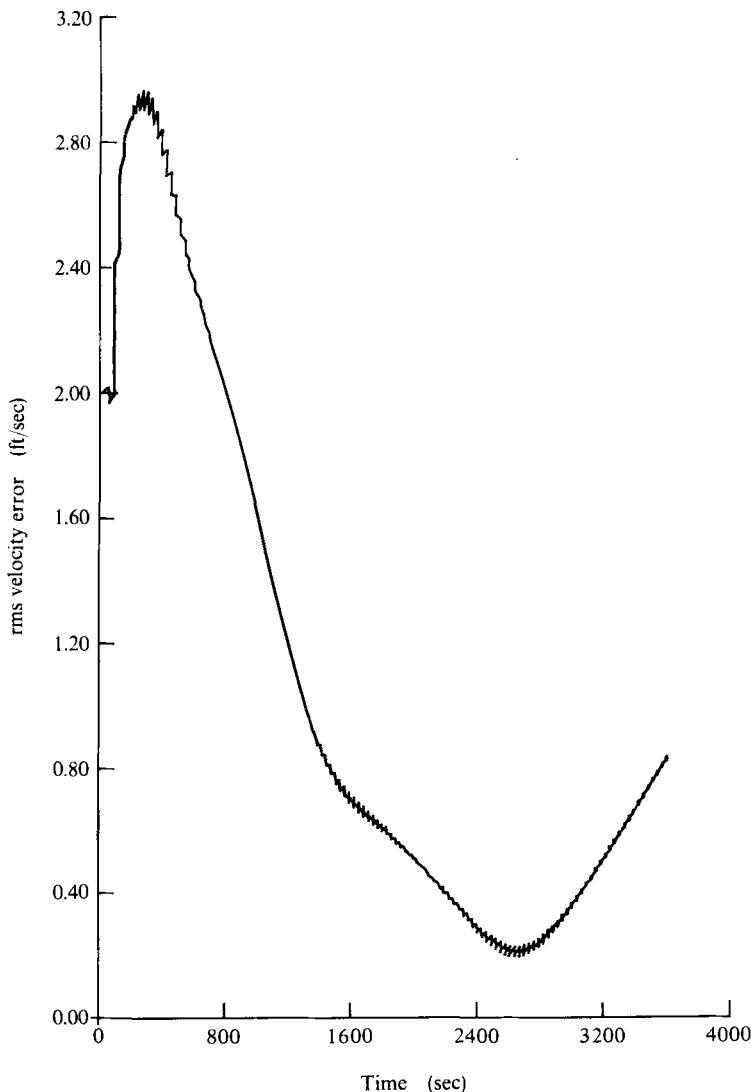
FIG. 6.27b   North rms velocity error of 10-state simplified filter.

• cycle time (and through it, the time to perform a load, store, add, multiply, divide, etc.);

• memory size and access;

• wordlength (Is rounding to nearest number or truncating least significant bits used, i.e., are wordlength errors symmetric or biased? Are square root forms or other means of enhancing numerical precision and stability warranted?);
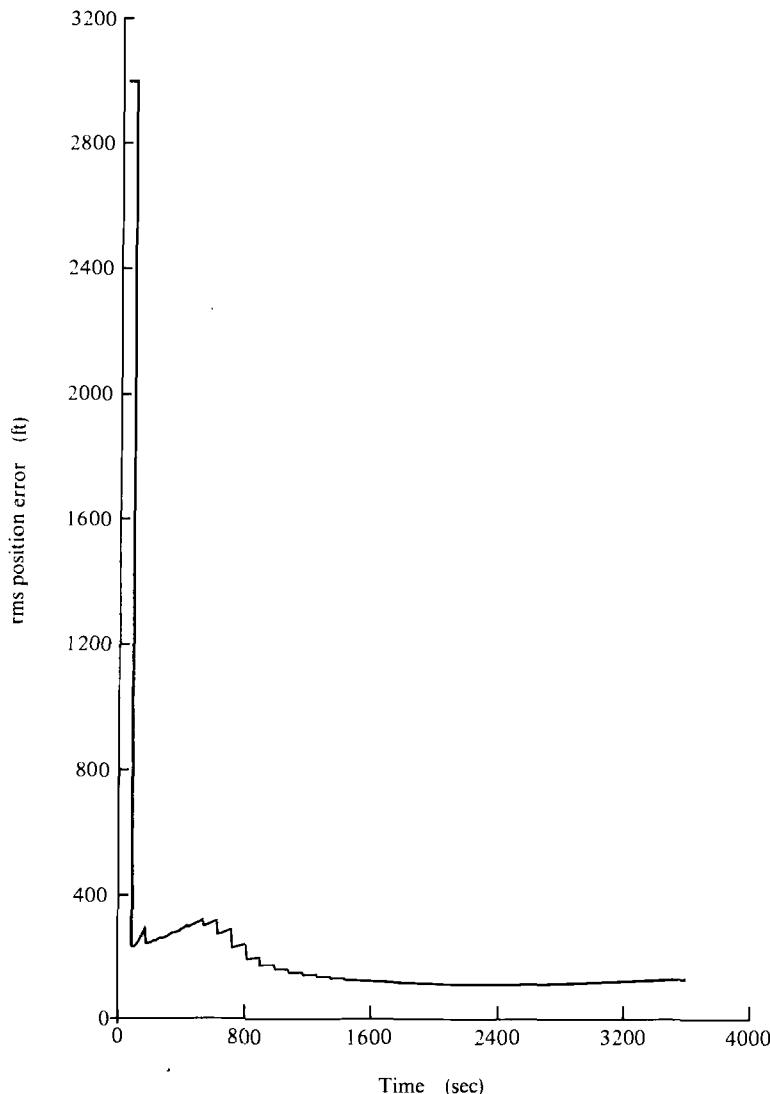
FIG. 6.28a   North rms position error of 15-state filter with 90-sec update period.

- readout A/D quantizations;
- instruction set;
- calculation capability;
- arithmetic type (floating or fixed point?).

In many applications, the filter will be one of many algorithms processed by a central digital computer. Consequently, the designer must plan real time
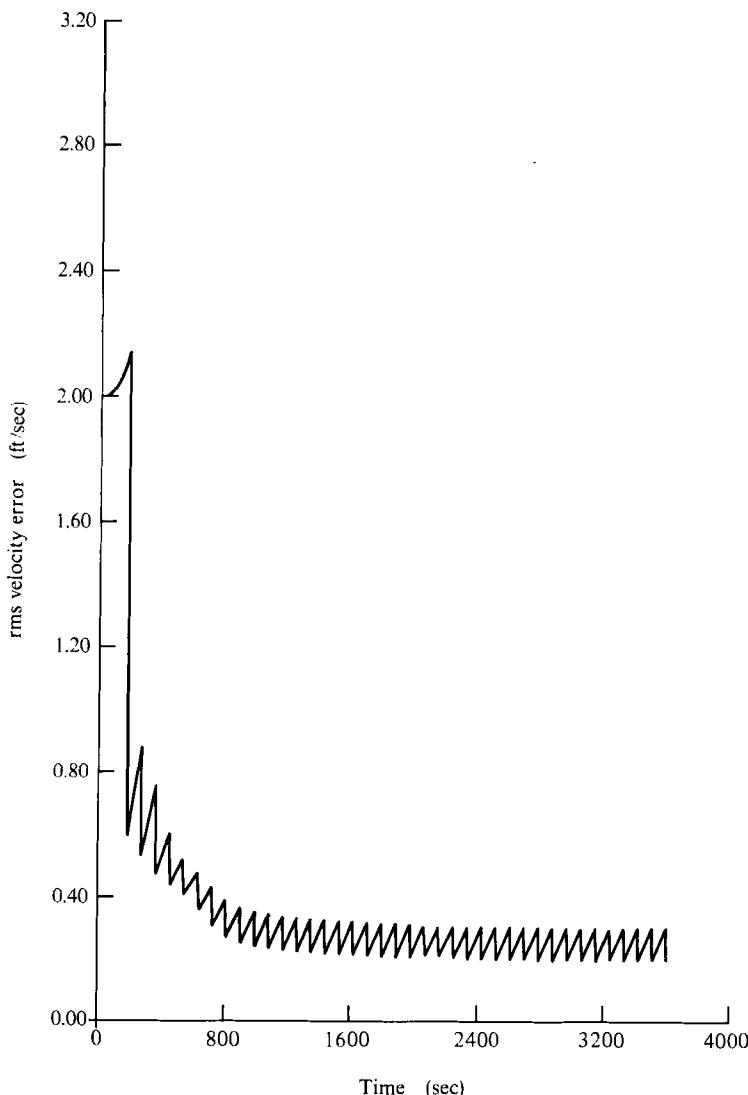
FIG. 6.28b   North rms velocity error of 15-state filter with 90-sec update period.

allocation and sequencing to be compatible with estimation accuracy specifications, sensor interfacing, and other software processing requirements. Often the iteration period of the filter is long compared to that of other jobs, such as a 30-sec period for an INS-aiding filter versus a 0.02-sec period for digital flight control inner loops. Therefore, the filter computations are typically performed "in the background" on a time-shared computer, admitting priority interrupts for more time-critical jobs.
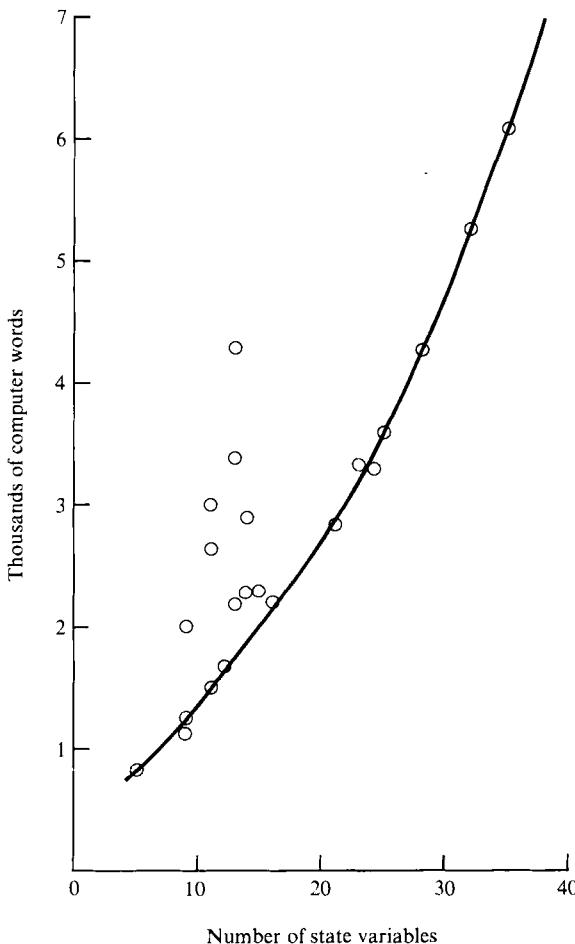
FIG. 6.29   Kalman filter computer memory requirements.

In the past, storage requirements have been a critical factor in the practicality of a filter design. Figure 6.29 plots the number of computer words required by actual navigation Kalman filters as a function of the number of state variables being estimated. These numbers account for both program instruction and permanent storage locations dedicated to the filter, but not erasable storage common to other programs. For instance, to implement a 16-state filter will require about 2200 words, about equally divided between instructions and permanent storage. In this figure, a curve has been drawn to indicate an approximate graph of this data for very efficient filters. Note that it does not exhibit a proportionality between required storage and the square of the state dimension as might be predicted theoretically (for $n \gg m$, the number of words required is predicted to be dominated by a term equal to $2.5n^2$). This is caused

by the fact that, as the state dimension of a filter grows, the defining matrices typically become considerably more sparse. Nevertheless, because computer memories are becoming dramatically less expensive, future filter designs may not focus so much attention on storage aspects.

Similarly, the number of computations can be calculated as a function of state variables used. For $n \gg m$, approximately $4n^3$ multiplications, and the same number of less time-consuming additions, would be needed for a single recursion of the filter. Again, this is not totally realized in practice due to increasing matrix sparsity with higher state dimension.

The practical constraints imposed by the computer do dictate a design philosophy of generating as simple a filter as possible that will meet performance specifications. Means of reducing complexity that are generally exploited are

- reducing state dimension while maintaining dominant facets of system behavior;
- neglecting terms, decoupling, and other model simplifications;
- canonical state space;
- exploiting symmetry (computing only lower triangular form of symmetric matrices);
- precomputations;
- approximations, as curve-fitted or constant gains;
- long sample periods, possibly using measurement prefiltering to make this feasible;
- iterative scalar measurement updating;
- removal of double precision requirements by use of a "square root filter" implementation (discussed in Chapter 7);
- approximation of stored constants as powers of two so that multiplications are replaced by simple shift operations; and
- efficient, rather than straightforward, programming of the algorithm such as replacing

$$\mathbf{P}^+ = \mathbf{P}^- - \mathbf{P}^- \mathbf{H}^T (\mathbf{H} \mathbf{P}^- \mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H} \mathbf{P}$$

by

$$\mathbf{A} = \mathbf{P}^- \mathbf{H}^T, \quad \mathbf{B} = \mathbf{H}\mathbf{A} + \mathbf{R}, \quad \mathbf{C} = \mathbf{B}^{-1} \quad \mathbf{P}^+ = \mathbf{P}^- - \mathbf{A}\mathbf{C}\mathbf{A}^T$$

Another practical consideration is the means of performing the estimate time propagations numerically. We will concentrate on the covariance time propagation. One method is to integrate the equation

$$\dot{\mathbf{P}}(t/t_{i-1}) = \mathbf{F}(t)\mathbf{P}(t/t_{i-1}) + \mathbf{P}(t/t_{i-1})\mathbf{F}^T(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^T(t) \qquad (6\text{-}110)$$

forward to time t_i from $\mathbf{P}(t_{i-1}/t_{i-1}) = \mathbf{P}(t_{i-1}^+)$. If we let Δt represent $[t_i - t_{i-1}]$, then simple Euler integration yields

$$\mathbf{P}(t_i^-) \overset{c}{=} \mathbf{P}(t_{i-1}^+) + [\dot{\mathbf{P}}(t_{i-1}/t_{i-1})] \Delta t \qquad (6\text{-}111)$$

where $\stackrel{c}{=}$ means computationally equivalent to. Integration accuracy is improved if the derivative in this relation were evaluated at the midpoint of the interval instead of the beginning:

$$\mathbf{P}(t_i^-) \stackrel{c}{=} \mathbf{P}(t_{i-1}^+) + \left[\dot{\mathbf{P}}(\tfrac{1}{2}t_{i-1} + \tfrac{1}{2}t_i/t_{i-1})\right]\Delta t \qquad (6\text{-}112\text{a})$$

$$\cong \mathbf{P}(t_{i-1}^+) + \tfrac{1}{2}\left[\dot{\mathbf{P}}(t_{i-1}/t_{i-1}) + \dot{\mathbf{P}}(t_i/t_{i-1})\right]\Delta t \qquad (6\text{-}112\text{b})$$

If neither of these sets of relations yield adequate performance, either the sample period can be subdivided and first order integration applied repeatedly to each step along the subintervals, or a higher order integration technique such as fourth order Runge–Kutta can be employed.

Another means of accomplishing the time propagation is through

$$\mathbf{P}(t_i^-) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\mathbf{\Phi}^{\mathrm{T}}(t_i, t_{i-1}) + \mathbf{Q}_{\mathrm{d}}(t_{i-1}) \qquad (6\text{-}113\text{a})$$

$$\mathbf{Q}_{\mathrm{d}}(t_{i-1}) = \int_{t_{i-1}}^{t_i} \mathbf{\Phi}(t_i, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^{\mathrm{T}}(\tau)\mathbf{\Phi}^{\mathrm{T}}(t_i, \tau)\,d\tau \qquad (6\text{-}113\text{b})$$

Chapter 4 indicated a first order approximation for $\mathbf{\Phi}(t_i, t_{i-1})$ and $\mathbf{Q}_{\mathrm{d}}(t_{i-1})$ as (see 4-132):

$$\mathbf{\Phi}(t_i, t_{i-1}) \stackrel{c}{=} \mathbf{I} + \mathbf{F}(t_{i-1})\Delta t \qquad (6\text{-}114\text{a})$$

$$\mathbf{Q}_{\mathrm{d}}(t_{i-1}) \stackrel{c}{=} \mathbf{G}(t_{i-1})\mathbf{Q}(t_{i-1})\mathbf{G}^{\mathrm{T}}(t_{i-1})\Delta t \qquad (6\text{-}114\text{b})$$

These numerical approximations are similarly improved if the matrices are evaluated at the midpoint of the interval instead of the beginning, or approximated by the average of their values at both ends of the interval. Moreover, subdividing the sample period and applying such relations repeatedly also will improve precision.

If the system model is time invariant, or has time variations that are slow enough to represent $\mathbf{F}$ adequately as constant over a single filter sample period, then the state transition matrix can be approximated by a truncated matrix exponential:

$$\mathbf{\Phi}(t_i, t_{i-1}) \stackrel{c}{=} \sum_{k=0}^{N} \frac{1}{k!}\left[\mathbf{F}(t_{i-1})\right]^k \Delta t^k \qquad (6\text{-}115)$$

where N might be chosen as two. Again, subpartitioning into J portions can be used to advantage to obtain

$$\mathbf{\Phi}(t_i, t_{i-1}) = \prod_{j=0}^{J-1} \mathbf{\Phi}(t_{i-1} + (j+1)\Delta t/J, \; t_{i-1} + j\Delta t/J) \qquad (6\text{-}116)$$

where each element in the product is calculated as in (6-115). A first order approximation to such a result would be

$$\mathbf{\Phi}(t_i, t_{i-1}) \stackrel{c}{=} \mathbf{I} + \sum_{j=0}^{J-1} \left[\mathbf{F}(t_{i-1} + j\Delta t/J)\right]\left[\Delta t/J\right] \qquad (6\text{-}117)$$

Other methods of evaluation were discussed in Problems 15–17 of Chapter 2.

If the first order approximation to $\mathbf{Q}_d(t_{i-1})$ given in (6-114b) is inadequate, other methods can be applied. First of all, the $\dot{\mathbf{Q}}(t, t_{i-1})$ equation to be integrated to yield $\mathbf{Q}_d(t_{i-1})$, (4-129c), is identical in form to (6-110), with the only difference being the initial condition. Therefore, the methods applicable to (6-110) are also usable here. Second, trapezoidal integration of (6-113b) yields

$$\mathbf{Q}_d(t_{i-1}) \stackrel{c}{=} \tfrac{1}{2}[\boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{G}(t_{i-1})\mathbf{Q}(t_{i-1})\mathbf{G}^T(t_{i-1})\boldsymbol{\Phi}^T(t_i, t_{i-1})$$
$$+ \mathbf{G}(t_i)\mathbf{Q}(t_i)\mathbf{G}^T(t_i)]\,\Delta t \qquad (6\text{-}118)$$

This form is especially attractive since it replaces having to know $\boldsymbol{\Phi}(t_i, \tau)$ for all τ with evaluating only $\boldsymbol{\Phi}(t_i, t_{i-1})$ as discussed previously; it is in fact the method used in the F-111 navigation filter [32]. If $\mathbf{GQG}^T$ is constant, then (6-114a) can be substituted into this expression and terms rearranged to yield another approximation as

$$\mathbf{Q}_d(t_{i-1}) \stackrel{c}{=} \tfrac{1}{2}[\boldsymbol{\Phi}(t_i, t_{i-1})\mathbf{GQG}^T + \mathbf{GQG}^T\boldsymbol{\Phi}^T(t_i, t_{i-1})]\,\Delta t \qquad (6\text{-}119)$$

which was used in a LORAN-inertial navigation filter for the Army [26]. Higher order integration techniques are possible, but are seldom necessary.

Residual monitoring as described in Section 5.4 is typically exploited in operational filters, at least for reasonableness checking and discarding of "bad" data points for which the residual magnitude exceeds some specified multiple of the computed standard deviation. Current designs embody more sophistication, using the residual monitoring for sensor failure declaration or adaptive purposes.

If a human operator is to provide measurements to the filter, the man–machine interface is a unique problem. Not only is update timing an issue, but so is the operator's confidence in the filter's perceived performance. For example, in the operation of the F-111 navigation filter, the operator can key in landmark position data. The accuracy ascribed to such data through associated diagonal elements in the $\mathbf{R}$ matrix was such that the Kalman gain was low and the navigation system updated its calculated position by only a small fraction of the residual. This dissatisfied navigators to the point where they attempted to "compensate" with fictitious position data or to override the filter function if they believed their own data to be precise [32]. To circumvent this lack of confidence, designers of a more recent navigation filter have incorporated the ability to "tell" the filter that the landmark information keyed in is thought to be poor, average, or very good: the navigator inputs both data and one of three predetermined R values to set the filter gain accordingly [1].

6.12 SUMMARY

This chapter has developed the means of exploiting the Kalman filter derived in the previous chapter, converting it from a result of mathematical optimization theory to a useful and flexible engineering tool. As has been

emphasized throughout the discussion, there are many possible filter designs for any given application. Physical insight and engineering judgment are essential to proposing viable designs, and then a thorough and accurate performance/loading tradeoff analysis must be conducted to generate a superior filter algorithm. A systematic approach to this iterative design procedure has been presented, one which minimizes the risk of erroneous decisions and major design modifications late in the filter development process, and one which maximizes the probability of an efficient, practical design that meets or surpasses all performance specifications. Because of constraints imposed by the computer and overall system of which the filter is an integral part, the design philosophy to be adopted is not to achieve the best possible performance at any cost, but to develop the simplest design that ensures attainment of performance goals.

Examples have concentrated upon the application of Kalman filtering to aiding inertial navigation systems. This has been an area benefited significantly by the Kalman filter, for two fundamental reasons. First of all, the error characteristics of an INS and other navigation aids complement each other well: the INS provides accurate high frequency data while the other sources supply good low frequency information. Thus, there is much benefit to be derived from combining such data through an optimal estimation algorithm. Perhaps more importantly, though, very adequate error models in the form of linear system representations driven by white Gaussian noise have been developed and validated for this application area, which is precisely the form required by the Kalman filter. As adequate models are developed in other areas, optimal estimation theory will become more widely exploited. Progressively the fundamental application of filtering theory will mature to a point where the systematic design procedure presented in this chapter can glean out the fullest potential of "optimal" filtering in the form of an efficient operational data processing algorithm.

REFERENCES

1. Bergeson, J., and Blahut, K., "B-1 Navigation System Mechanization," Tech. Rep. D229-10336-1, The Boeing Co., Seattle, Washington, May 1974.
2. Britting, K. R., Inertial Navigation System Analysis. Wiley, New York, 1971.
3. Brock, L. D., and Schmidt, G. T., "General Questions on Kalman Filtering in Navigation Systems," in Theory and Applications of Kalman Filtering (C. T. Leondes, ed.), AGARDograph No. 139. NATO Advisory Group for Aerospace Research and Development, London, February 1970.
4. Broxmeyer, C., Inertial Navigation Systems. McGraw-Hill, New York, 1964.
5. Butler, R. R., and Rhue, G. T., "Kalman Filter Design for an Inertial Navigation System Aided by Non-Synchronous Navigation Satellite Constellations," Masters thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, March 1974.
6. Chalk, C. R. et al., "Background Information and User Guide for MIL F-8785B(ASG) Entitled Military Specification—Flying Qualities of Piloted Airplanes," AFFDL TR-69-72, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, August 1969.

7. Clark, R. R., "Performance Sensitivity Analysis of a Kalman Filter Using Adjoint Functions," NAFI TR-1767, Naval Avionics Facility, Indianapolis, Indiana, February 1972.

8. D'Appolito, J. A., "The Evaluation of Kalman Filter Designs for Multisensor Integrated Navigation Systems," AFAL-TR-70-271, The Analytic Sciences Corp., Reading, Massachusetts, January 1971.

9. D'Appolito, J. A., and Hutchinson, C. E., "Low Sensitivity Filters for State Estimation in the Presence of Large Parameter Uncertainties," *IEEE Trans. Automatic Control* **AC-14** (3), 310–312 (1969).

10. D'Appolito, J. A., and Roy, K. J., "Satellite/Inertial Navigation System Kalman Filter Design Study," Tech. Rep. AFAL-TR-71-178, Air Force Avionics Laboratory, Wright-Patterson AFB, Ohio, 1971.

11. Fagin, S. L., "Recursive Linear Regression Theory, Optimal Filter Theory and Error Analysis of Optimal Systems," *IEEE Internat. Convent. Record* 216–240 (1964).

12. Fitzgerald, R. J., "Divergence of the Kalman Filter," *IEEE Trans. Automatic Control* **AC-16** (6), 736–747 (1971).

13. Friedland, B., "On the Effect of Incorrect Gain in the Kalman Filter," *IEEE Trans. Automatic Control* **AC-12** (5), 610 (1967).

14. Friedland, B., "Treatment of Bias in Recursive Filtering," *IEEE Trans. Automatic Control* **AC-14** (4), 359–367 (1969).

15. Griffin, R. E., and Sage, A. P., "Large and Small Scale Sensitivity Analysis of Optimum Estimation Algorithms," *IEEE Trans. Automatic Control* **AC-13** (4), 320–329 (1968).

16. Griffin, R. E., and Sage, A. P., "Sensitivity Analysis of Discrete Filtering and Smoothing Algorithms," *AIAA J.* **7** (10), 1890–1897 (1969).

17. Hamilton, E. L., Chitwood, G., and Reeves, R. M., "An Efficient Covariance Analysis Computer Program Implementation," *Proc. Nat. Aerospace and Electronic. Conf., Dayton, Ohio* (May 1976).

18. Hamilton, E. L., Chitwood, G., and Reeves, R. M., "The General Covariance Analysis Program (GCAP): An Efficient Implementation of the Covariance Analysis Equations," Air Force Avionics Laboratory, Wright-Patterson AFB, Ohio, June 1976.

19. Hammett, J. E., "Evaluation of a Proposed INS Kalman Filter in a Dynamic Flight Environment," Masters thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, December 1974.

20. Heffes, H., "The Effect of Erroneous Models on the Kalman Filter Response," *IEEE Trans. Automatic Control* **AC-11** (3), 541–543 (1966).

21. Heller, W. G., "Models for Aided Inertial Navigation System Sensor Errors," TASC TR-312-3, The Analytic Sciences Corp., Reading, Massachusetts, February 1975.

22. Hollister, W. M., and Bansal, D. D., "Guidance and Control for V/STOL Aircraft—Final Report," DOT-TSC-5, Measurement Systems Laboratory RE-77, M.I.T., Cambridge, Massachusetts, November 1970.

23. Hollister, W. M., and Brayard, M. C., "Optimum Mixing of Inertial Navigator and Position Fix Data," Rep. RE-62, M.I.T. Measurement Systems Laboratory, Cambridge, Massachusetts., August 1969; also AIAA Paper No. 70–35 at the Eighth Aerospace Sciences Meeting, New York, January 1970.

24. Huddle, J. R., and Wismer, D. A., "Degradation of Linear Filter Performance Due to Modeling Error," *IEEE Trans. Automatic Control* **AC-13** (4), 421–423 (1968).

25. Kayton, M., and Fried, W. R., *Avionics Navigation Systems.* Wiley, New York, 1969.

26. Knight, J., Light, W., and Fisher, M., "Selected Approaches to Measurement Processing and Implementation in Kalman Filters," Tech. Rep. ECOM-3510, U.S. Army Electronics Command, Fort Monmouth, New Jersey, November 1971.

27. Kwakernaak, H., "Sensitivity Analysis of Discrete Kalman Filters," *Internat. J. Control* **12**, 657–669, (1970).

28. Leondes, C. T. (ed.), *Theory and Applications of Kalman Filtering*, AGARDograph No. 139. NATO Advisory Group for Aerospace Research and Development, London, February 1970.

29. Maybeck, P. S., "The Kalman Filter—An Introduction for Potential Users," TM-72-3, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, June 1972.

30. Maybeck, P. S., "Filter Design for a TACAN-Aided Baro-Inertial System with ILS Smoothing Capability," AFFDL-TM-74-52, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, January 1974.

31. Maybeck, P. S., "Applied Optimal Estimation: Kalman Filter Design and Implementation," notes for a short course presented by the Air Force Institute of Technology, Wright-Patterson AFB, Ohio, semiannually since December 1974.

32. Maybeck, P. S., "Review of F-111 Kalman Filter Meeting at Sacramento ALC, 26 July 1976," correspondence, July 1976; also "USAF Tech. Order IF-111D-2-5-1-1" (anon.); "Customer Training Document on F-111 Avionics," (General Dynamics, Fort Worth); "System Program Description Document for the F-111D General Navigation and Weapon Delivery Computers," (FZE-12-8078A, USAF, June 1976); "Comments and Questions on the F-111D Kalman Filter" (Maybeck, P. S., June 1976); "Discussion of Capt. Maybeck's Comments and Questions on the F-111D Kalman Filter" (DeVries, T. W., Autonetics Division of Rockwell International, July 1976).

33. Maybeck, P. S., "Analysis of a Kalman Filter for a Strapdown Inertial/Radiometric Area Correlator Guidance System," *Proc. IEEE Nat. Aerospace and Electron Conf., Dayton, Ohio* (May 1977).

34. Maybeck, P. S., "Performance Analysis of a Particularly Simple Kalman Filter," *AIAA J. Guidance and Control* **1** (6), 391–396 (1978).

35. Mendel, J. M., "Computational Requirements for a Discrete Kalman Filter," *IEEE Trans. Automatic Control* **AC-16** (6), 748–758, (1971).

36. Nash, R. A., D'Appolito, J. A., and Roy, K. J., "Error Analysis of Hybrid Inertial Navigation Systems," *Proc. AIAA Guidance and Control Conf.* Paper No. 72-848, Stanford, California, (August 1972).

37. Nash, R. A., and Tuteur, F. B., "The Effect of Uncertainties in the Noise Covariance Matrices on the Maximum Likelihood Estimate of a Vector," *IEEE Trans. Automatic Control* **AC-13** (1), 86–88 (1968).

38. Neal, S. R., "Linear Estimation in the Presence of Errors in Assumed Plant Dynamics," *IEEE Trans. Automatic Control* **AC-12** (5), 592–594 (1967).

39. Nishimura, T., "On the A Priori Information in Sequential Estimation Problems," *IEEE Trans. Automatic Control* **AC-11** (2), 197–204 (1966); correction to and extension of, **AC-12** (1), 123 (1967).

40. Pinson, J. C., "Inertial Guidance for Cruise Vehicles," *in Guidance and Control of Aerospace Vehicles* (C. T. Leondes, ed.). McGraw-Hill, New York, 1963.

41. Pitman, G. R. (ed.), *Inertial Guidance*. Wiley, New York, 1962.

42. Potter, J. E., and Vander Velde, W. E., "Optimum Mixing of Gyroscope and Star Tracking Data," Rep. RE-26, MIT Experimental Astronomy Laboratory, Cambridge, Massachusetts, 1968; also *AIAA J. Spacecraft and Rockets* **5** (5), 536–540 (1968).

43. Price, C. F., "An Analysis of the Divergence Problem in the Kalman Filter," *IEEE Trans. Automatic Control* **AC-13** (6), 699–702 (1968).

44. Schlee, F. H., Standish, C. J., and Toda, N. F., "Divergence in the Kalman Filter," *AIAA J.* **5** (6), 1114–1120 (1967).

45. Schmidt, G. T., "Aided and Optimal Systems," unpublished paper, M.I.T., Cambridge, Massachusetts, September 1968; also "M.I.T. Lectures on Kalman Filtering," C. S. Draper Rep. P-061, Cambridge, Massachusetts, 1974.

46. Schmidt, G. T., "Linear and Nonlinear Filtering Techniques," *in Control and Dynamic Systems* (C. T. Leondes, ed.), Vol. 12. Academic Press, New York, 1976.

47. Schmidt, G. T. (ed.), *Practical Aspects of Kalman Filtering Implementation*, AGARD-LS-82, NATO Advisory Group for Aerospace Research and Development, London, May 1976.

48. Schmidt, S. F., "Computational Techniques in Kalman Filtering," *in Theory and Applications of Kalman Filtering*, AGARDograph 139, NATO Advisory Group for Aerospace Research and Development, London, February 1970.

49. Simon, K. W., and Stubberud, A. R., "Reduced Order Kalman Filter," *Internat. J. Control* **10**, 501–509 (1969).

50. Sims, F. L., and Lainiotis, D. G., "Sensitivity Analysis of Discrete Kalman Filters," *Conf. Record Asilomar Conf. on Circuits and Systems, 2nd*, pp. 147–152 (1968).

51. Sorenson, H. W., "On the Error Behavior in Linear Minimum Variance Estimation Problems," *IEEE Trans. Automatic Control* **AC-12** (5), 557–562 (1967).

52. Widnall, W. S., and Grundy, P. A., "Inertial Navigation System Error Models," Intermetrics TR-03-73, Intermetrics, Inc., Cambridge, Massachusetts, May 1973.

PROBLEMS

6.1 *Combining Inertial and ILS Information* A substantial amount of work is now being conducted in the area of combining inertial and instrument landing system (ILS) signals by means of a Kalman filter. An ILS is composed of two beams, one oscillating in the vertical plane (which provides the glide-slope measurement) and another in the horizontal plane (to provide the localizer measurement). These oscillate in restricted arcs about the nominal approach path to the runway.

It is proposed that, once the localizer beam and glide-slope beam are captured, the ILS and inertial data be combined optimally. In effect, the inertial system is used to smooth the ILS data before presenting the approach information to the pilot.

It is assumed that the aircraft will have achieved a location and velocity "close" to a nominal trajectory to the runway. The inertial system (possibly aided during the preceding cruise segment of flight) will provide the necessary initial conditions. Furthermore, initial offsets in the stable member alignment from vertical, accumulated velocity errors, and the effects of gyro drift during the ILS/INS mode will have negligible effects and can be ignored.

For category II (restricted ceiling and visual range) landings, ILS beam errors and onboard receiver errors are expected to cause deviations from the ideal trajectory according to the following table of path angle *peak* deviations (outer and middle markers are located in line with the runway along the approach trajectory and threshold refers to the end of the runway).

Location	Localizer deviation	Glide-slope deviation
Beyond outer marker	$\pm 0.4°$	$\pm 0.1°$
Outer marker to middle marker	$\pm 0.4°$ decreasing to $\pm 0.07°$	$\pm 0.1°$ decreasing to $\pm 0.06°$
Middle marker to threshold	$\pm 0.07°$	$\pm 0.06°$

(a) How would you incorporate this data (a typical specification) into the format required in the Kalman filter formulation?

The localizer measurement at time instant t_i is modeled by

$$\mathsf{L}_{\mathrm{ILS}}(t_i) = \mathsf{L}_{\mathrm{true}}(t_i) + \delta\mathsf{L}_{\mathrm{ILS}}(t_i) - \mathsf{v}_1(t_i) \tag{1}$$

and the glide-slope measurement is

$$\mathsf{S}_{\mathrm{ILS}}(t_i) = \mathsf{S}_{\mathrm{true}}(t_i) + \delta\mathsf{S}_{\mathrm{ILS}}(t_i) - \mathsf{v}_2(t_i) \tag{2}$$

Here v_1 and v_2 are white Gaussian noises with zero mean and variances $\sigma_{v_1}^2$ and $\sigma_{v_2}^2$, respectively (and they are assumed to be uncorrelated). However, empirical data revealed that a simple white noise added to the true values did not model the true errors sufficiently, so δL_{ILS} and δS_{ILS} were added. These are modeled as zero-mean, exponentially time-correlated Gaussian processes with

$$E\{\delta L_{ILS}(t_i)\,\delta L_{ILS}(t_j)\} = \sigma_L^2 e^{-|t_i - t_j|/T} \quad \cdot \tag{3}$$

$$E\{\delta S_{ILS}(t_i)\,\delta S_{ILS}(t_j)\} = \sigma_S^2 e^{-|t_i - t_j|/T} \tag{4}$$

where T is the correlation time common to both. It can be shown that a signal with such a time correlation can be produced by passing a zero-mean, white Gaussian noise through a first order lag. In sampled data form, the "shaping filter" equations are

$$\delta L_{ILS}(t_i) = e^{-\Delta t/T}\,\delta L_{ILS}(t_{i-1}) + w_{d1}(t_{i-1}) \tag{5}$$

$$\delta S_{ILS}(t_i) = e^{-\Delta t/T}\,\delta S_{ILS}(t_{i-1}) + w_{d2}(t_{i-1}) \tag{6}$$

where Δt is the time between samples, $t_i - t_{i-1}$.

(b)   Determine the appropriate value of the variances of w_{d1} and w_{d2} noises such that the steady state outputs of the "shaping filters" satisfy Eqs. (3) and (4).

The inertial system provides predictions of the deflections from the desired glide-slope path, based upon the current latitude, longitude, and altitude as calculated by the INS. These inertial predictions are expressible as

$$L_{INS}(t_i) = L_{true}(t_i) + \delta L_{INS}(t_i) \tag{7}$$

$$S_{INS}(t_i) = S_{true}(t_i) + \delta S_{INS}(t_i) \tag{8}$$

Now, if x, y, and z are, respectively, the east, north, and upward directions, then the above equations can be linearized about a nominal approach trajectory.

(c)   Express $\delta L_{INS}(t_i)$ and $\delta S_{INS}(t_i)$ in terms of $\delta x(t_i)$, $\delta y(t_i)$, $\delta z(t_i)$, the INS errors. To do this, it is suggested that a geometrical picture be drawn to display true and INS-indicated x, y, z, L, and S. Knowledge of the location and heading of the runway can be assumed to be prestored information in the computer.

As in the simple example in Section 6.4, the INS can be modeled as (for periods of time as short as those of interest for landing) a double integrator of acceleration:

$$\delta x(t_i) = \delta x(t_{i-1}) + (\Delta t)\,\delta v_x(t_{i-1}) \tag{9}$$

$$\delta v_x(t_i) = \delta v_x(t_{i-1}) + w_{dx}(t_{i-1}) \tag{10}$$

with w_{dx} a white Gaussian zero-mean noise sequence with variance Q_x. Similar equations would then exist for the y and z directions as well.

(d)   Write out the discrete-time equations that describe the error state space description of this system model. There are eight state variables, but an eight-dimensional state vector is not necessary: decoupling is possible. Decouple the equations into the appropriate number of sets of equations of the form

$$\mathbf{x}(t_i) = \mathbf{\Phi}\mathbf{x}(t_{i-1}) + \mathbf{G}_d\mathbf{w}_d(t_{i-1}), \qquad \mathbf{z}(t_i) = \mathbf{H}\mathbf{x}(t_i) + \mathbf{v}(t_i)$$

These decoupled equations define the state space model for separate Kalman filters.

(e)   Write out the Kalman filter equations. Note that the time propagation is simplified if an indirect feedback filter is implemented.

(f)   Consider the possibility of nearly optimal simplified gains for this problem. How might it be done?

6.2 The noise random variable $v(t_i)$ in (6-52) is a Gaussian noise meant to model the effects of quantization on the measurement at time t_i, and this problem establishes the appropriate description of such a noise. Suppose that an analog-to-digital converter or some other quantizer operates with a quantization level of Δ.

(a) If it uses "rounding," the device will convert any input $r(t_i)$ in the range

$$[k\Delta - \tfrac{1}{2}\Delta] \leq r(t_i) < [k\Delta + \tfrac{1}{2}\Delta]$$

into an output $y(t_i) = k\Delta$. If all input values are equally likely, a good model of the error due to quantization is a uniformly distributed random variable. Completely describe its density function, mean, and variance.

(b) If the quantizer uses "truncation," converting any input $r(t_i)$,

$$[k\Delta] \leq r(t_i) < [k\Delta + \Delta]$$

into an output $y(t_i) = k\Delta$, how does this change the results in (a)?

(c) Gaussian random variables can be used to approximate the uniform variables just proposed. An equivalent mean can be set and either the variance can be set equal to the variances found in (a) and (b), or the 3σ value set equal to half the range (i.e., $\Delta/2$). Generate these approximations and plot the results.
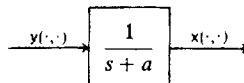
6.3 If numerical problems arise in a filter due to a wide dynamic range of values of interest (as covariance matrix eigenvalues for example), rescaling of variables is often conducted to reduce this range.

(a) Let a three-dimensional filter be described in terms of variables x_1, x_2, and x_3. Convert this to a filter expressed in terms of states cx_1, x_2, and x_3, where c is a scale factor. Describe the new filter defining quantities in terms of the original $\mathbf{F}$, $\mathbf{B}_d$ (let $r = 2$), $\mathbf{G}_d\mathbf{Q}_d\mathbf{G}_d{}^T$, $\hat{\mathbf{x}}_0$, $\mathbf{P}_0$, $\mathbf{H}$ (let $m = 2$), and $\mathbf{R}$. First do this by writing the scalar equations explicitly, then by means of an appropriate similarity transformation.

(b) Generalize this to describe the filter in terms of variables (c_1x_1), (c_2x_2) and (c_3x_3), with c_1, c_2, and c_3 chosen scale factors.

6.4 Show that $\mathbf{x}_a(\cdot,\cdot)$ defined in (6-81) is a Gaussian process.

6.5 A good model for a given system is that it is composed of a first order lag, with transfer function $1/(s + a)$:

where $y(\cdot,\cdot)$ can be modeled as a stationary, exponentially time-correlated zero-mean Gaussian noise with correlation time T:

$$E\{y(t)y(t + \tau)\} = 5e^{-|\tau|/T}$$

Discrete-time measurements are available as

$$z(t_i) = x(t_i) + v(t_i)$$

where $v(\cdot,\cdot)$ is a zero-mean, white Gaussian noise with variance

$$E\{v(t_i)^2\} = 2$$

The system starts at rest.

(a) Generate the "truth" model for this problem, explicitly depicting $\mathbf{F}_t$, $\mathbf{G}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$, $\mathbf{R}_t$, $\mathbf{P}_{t0}$, and $\mathbf{C}_t$ [assuming $x(\cdot,\cdot)$ to be the quantity of basic interest].

(b) To build an efficient Kalman filter, you might consider simplifying the model by replacing the time-correlated noise with a white noise of appropriate strength in the model upon which the filter is based. What is the appropriate strength of the white noise? What relationships between sample period Δt, correlation time T, and system parameter a would yield confidence in this being an adequate model?

(c) Specify the Kalman filter equations explicitly for propagating and updating the optimal estimate of the system state, depicting and justifying choices for $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{P}_0$, and $\mathbf{C}$.

(d) Explicitly write out the equations that would be used for a covariance performance analysis (sensitivity analysis) of the resulting filter operating in an environment defined by the "truth" model of part (a).

6.6 In the previous problem, if the strength of white noise in the reduced-order model were chosen so as to duplicate the low frequency power spectral density value of the original noise, then $Q = 10T$. Another means of choosing Q would be such that the steady state variance of $x(\cdot,\cdot)$ generated by the reduced order model is equivalent to that of the original model. Show that this yields $Q = [10T]/[1 + aT]$. Is this significantly different?

6.7 Consider an estimator algorithm identical in structure to a Kalman filter but with a gain matrix $\bar{\mathbf{K}}(t_i)$ different from the Kalman gain $\mathbf{K}(t_i)$ for each time t_i. Show that if the estimate error covariance before measurement incorporation is $\mathbf{P}(t_i^-)$, then the covariance of the error committed by the estimate after measurement incorporation is

$$\mathbf{P}(t_i^+) = [\mathbf{I} - \bar{\mathbf{K}}(t_i)\mathbf{H}(t_i)]\mathbf{P}(t_i^-)[\mathbf{I} - \bar{\mathbf{K}}(t_i)\mathbf{H}(t_i)]^{\mathrm{T}} + \bar{\mathbf{K}}(t_i)\mathbf{R}(t_i)\bar{\mathbf{K}}^{\mathrm{T}}(t_i)$$

by first writing

$$\hat{\mathbf{x}}(t_i^+) = [\mathbf{I} - \bar{\mathbf{K}}(t_i)\mathbf{H}(t_i)]\hat{\mathbf{x}}(t_i^-) + \bar{\mathbf{K}}(t_i)\mathbf{z}(t_i)$$

Thus the Joseph form update generalizes to this case, whereas other forms do not.

Show that the analogous result for a continuous-time, continuous-measurement linear filter is that $\mathbf{P}(t)$ satisfies

$$\dot{\mathbf{P}}(t) = [\mathbf{F}(t) - \bar{\mathbf{K}}(t)\mathbf{H}(t)]\mathbf{P}(t) + \mathbf{P}(t)[\mathbf{F}(t) - \bar{\mathbf{K}}(t)\mathbf{H}(t)]^{\mathrm{T}} + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^{\mathrm{T}}(t) + \bar{\mathbf{K}}(t)\mathbf{R}(t)\bar{\mathbf{K}}^{\mathrm{T}}(t).$$

6.8 What modifications have to be made to the algorithm of Section 6.8 to evaluate approximated gains rather than the true Kalman gain in the filter? What if you wished to evaluate some general linear predictor–corrector filter forms?

6.9 Both Monte Carlo and covariance analysis relationships can be developed analogous to those in Section 6.8 by replacing the augmented state vector process $\mathbf{x}_a(\cdot,\cdot)$ defined in (6-81) by an augmented vector

$$\mathbf{x}_a'(\cdot,\cdot) = \left[\begin{array}{c} \mathbf{e}_t'(\cdot,\cdot) \\ \hline \hat{\mathbf{x}}(\cdot/t_{i-1},\cdot) \end{array}\right]$$

where

$$\mathbf{e}_t'(\cdot,\cdot) \triangleq \mathbf{x}_t(\cdot,\cdot) - \mathbf{T}\hat{\mathbf{x}}(\cdot/t_{i-1},\cdot)$$

and $\mathbf{T}$ is an (n_t)-by-n matrix relating the n filter states to the n_t states of the truth model. In many cases, $\mathbf{T}$ is of the form

$$\mathbf{T} = \left[\begin{array}{c} \mathbf{I} \\ \mathbf{0} \end{array}\right] \begin{array}{l} \} \ n \ \text{rows} \\ \} \ (n_t - n) \ \text{rows} \end{array}$$

Then it is the process $\mathbf{e}_t'(\cdot,\cdot)$ directly, or a covariance description of this process, which is of interest as an analysis output, comparable to $\mathbf{e}_t(\cdot,\cdot)$ of (6-95) or (6-104). Develop the appropriate Monte Carlo state relations and covariance matrix relations for this formulation.

6.10   Error ellipsoids are typically employed to obtain a geometrical interpretation of an error covariance matrix for a Gaussian random vector. For an n-dimensional, zero-mean error $\mathbf{x}$, a family of ellipsoids having surfaces of constant probability density can be defined through

$$\boldsymbol{\xi}^T \mathbf{P}^{-1} \boldsymbol{\xi} = k$$

where k is some arbitrary constant. The "error ellipsoid corresponding to probability $\mathscr{P}$" is the particular ellipsoid for which the probability that the error $\mathbf{x}(\omega_j) = x$ lies inside that ellipsoid is $\mathscr{P}$. The principal axes of that error ellipsoid can be described by the n vectors $\sqrt{k\lambda_i}\mathbf{e}_i$ where the λ_i's are the eigenvalues of $\mathbf{P}$ and the $\mathbf{e}_i$'s are the corresponding eigenvectors.

The following table presents the values of k and $\sqrt{k}$ for $\mathscr{P} = 0.6826$ and 0.9974 (corresponding to 1σ and 3σ ellipses for the scalar case) for error vector dimensions $n = 1,2,3,4,6$. These are commonly chosen error ellipsoids for analysis purposes. Also presented is the ratio $[k(\mathscr{P} = 0.9974)]^{1/2}/[k(\mathscr{P} = 0.6826)]^{1/2}$, used as a multiplicative factor on the axis dimensions of the first case ellipsoid to generate those of the second.

Verify these results.

Error vector dimension n	$\mathscr{P} = 0.6826$		$\mathscr{P} = 0.9974$		$\dfrac{[k(\mathscr{P} = 0.9974)]^{1/2}}{[k(\mathscr{P} = 0.6826)]^{1/2}}$
	k	$\sqrt{k}$	k	$\sqrt{k}$	
1	1.000	1.000	9.000	3.000	3.00
2	2.296	1.515	11.820	3.438	2.27
3	3.527	1.878	14.157	3.763	2.00
4	4.720	2.172	16.251	4.031	1.86
6	7.038	2.653	20.062	4.479	1.69

6.11   Suppose that a two-dimensional Gaussian random vector expressed in principal coordinates (i.e., with diagonal covariance matrix, with σ_x^2 and σ_y^2 as diagonal terms) can be used to describe a planar distribution of interest, such as the location of the splashdown landing site for a returning manned spacecraft. A performance description often employed is the CEP, the circular error probability, the radius of the *circle* that contains 50% of the realizations of the random vector.

(a)   Show that, if $\sigma_x = \sigma_y = \sigma$ and the circle is centered at the mean vector $[m_x, m_y]^T$, that the CEP is 1.177σ.

(b)   Let $\sigma_x > \sigma_y$ and generate the integral relation to solve for CEP, assuming the circle to be centered at the mean vector. Two commonly used approximations to this result are

$$\widehat{\text{CEP}}_1 = 0.588[\sigma_x + \sigma_y], \qquad \widehat{\text{CEP}}_2 = 0.563\sigma_x + 0.614\sigma_y$$

Compute true and approximated CEP's for cases of $\sigma_x = \sigma_y$, $2\sigma_y$, $3\sigma_y$, and $4\sigma_y$. The error in the second approximation is less then 1°_0 to about $(\sigma_x/\sigma_y) = 3$, and less than 10°_0 to about $(\sigma_x/\sigma_y) = 10$, beyond which point the severe ellipticity of equiprobability loci makes CEP a poor means of performance description.

(c)   If the CEP circle is not centered at the mean value, is the CEP larger or smaller? Explain. Numerical integration is necessary for determining CEP in these cases.

6.12   This problem is meant to indicate the extreme care necessary in interpreting a certain means of conducting and graphically presenting error budget type information. Instead of analyzing the individual effects of N different error sources as done to generate Fig. 6.21, one might progressively add each source to the previous sources on N separate runs. Then a chart might be plotted as in Fig. 6.P1:

Let $N = 3$ and assume that the contribution to some rms error of interest by each separate source (as plotted in Fig. 6.21) can be denoted as σ_1, σ_2, and σ_3 due to sources numbered 1, 2, and 3, respectively.

(a) Let $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$ and plot both graphs. Which source yields the greatest effect on system performance?

(b) Let source 2 be incorporated first, then source 3, then source 1. Repeat part (a). Is there a fallacy in the relative importance of sources inferred from the progressive addition method?

(c) Let sources be added progressively in numerical order, and let $\sigma_1 = 1$, $\sigma_2 = 1.5$, and $\sigma_3 = 1.7$. What ordering of relative importance is suggested by the graphical methods just described?

6.13 Demonstrate the validity of the computational forms of (6-118) and (6-119).

6.14 A human navigator has erroneously keyed in the wrong position update information into an optimally aided inertial system employing a Kalman filter. He immediately keys in the correct data twice in succession, seeking to force the filter to pay significantly greater attention to the good data than the bad. Will this work the way he hopes?

Error due to source 1   Error due to sources 1 and 2   Etc.   Total error
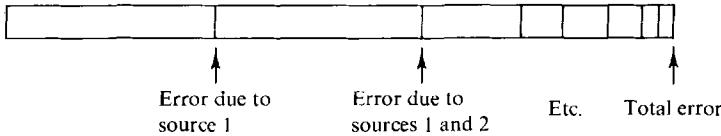
FIG. 6.P1   Graph of data for Problem 6.12.

Square root filtering

7.1 INTRODUCTION

The two previous chapters discussed the Kalman filter in substantial detail. Although this is the optimal solution to the filtering problem posed in Section 5.2 (with respect to essentially all viable optimality criteria), the algorithm itself is prone to serious numerical difficulties. As noted in Section 5.6, measurement updating of the covariance matrix requires rather long wordlength to maintain acceptable numerical accuracy: for onboard computers, double precision computations are usually required. In fact, without double precision arithmetic, these numerical characteristics can easily become the dominant error source corrupting estimation precision, and unfortunately an error source usually not included in designers' error budgets.

The difficulties encountered in converting a tuned Kalman filter on a long wordlength, large computer system used for engineering design to an effective algorithm on a smaller wordlength online computer are well documented [7, 22]. For instance, although it is theoretically impossible for the covariance matrix to have negative eigenvalues, such a condition can, and often does, result due to numerical computation using finite wordlength, especially when (1) the measurements are very accurate [eigenvalues of $R(t_i)$ are small relative to those of $P(t_i^-)$, this being accentuated by large eigenvalues in P_0] or (2) a linear combination of state vector components is known with great precision while other combinations are nearly unobservable (i.e., there is a large range of magnitudes of state covariance eigenvalues). Such a condition can lead to subsequent divergence or total failure of the recursion. On close inspection, even Kalman filters that maintain adequate estimation accuracy exhibit instances of negative covariance diagonal terms [7].

To circumvent these problems in numerics inherent to the Kalman filter algorithm, alternate recursion relationships [24] have been developed to propagate and update a state estimate and error covariance square root or inverse covariance square root instead of the covariance or its inverse themselves. Although equivalent algebraically to the conventional Kalman filter recursion, these *square root filters* exhibit improved numerical precision and stability, particularly in ill-conditioned problems (i.e., the cases described that yield erroneous results due to finite wordlength). The square root approach can yield twice the effective precision of the conventional filter in ill-conditioned problems. In other words, the same precision can be achieved with approximately half the wordlength. Moreover, this method is completely successful in maintaining the positive semidefiniteness of the error covariance.

These excellent numerical characteristics, combined with modest additional computation cycle time and memory storage requirements, make the square root filter approach a viable alternative to the conventional filter in many applications, especially when computer wordlength is limited or the estimation problem is ill conditioned. The formulation of the square root filter for the case of no dynamic noise is especially attractive because of its computational simplicity, and its outstanding numerical characteristics led to its implementation in the Apollo spacecraft navigation filters.

A number of practitioners have argued, with considerable logic, that square root filters should *always* be adopted in preference to the standard Kalman filter recursion, rather than switching to these more accurate and stable algorithms when and if numerical problems occur [7]. Even though Kalman algorithms can be made to work in most applications, by using double precision mathematics or scaling variables to reduce dynamic range or employing ad hoc modifications, numerics degrade performance from that achievable by numerically better conditioned recursions. Recent investigations tend to support an approach of designing and tuning an optimal filter by the methods of the two previous chapters, ignoring the errors caused by numerics, but then implementing the square root equivalent for online operation. Nevertheless, one can expect conventional Kalman algorithms to be applied rather extensively as well.

Section 7.2 introduces the concept of matrix square roots, and then Section 7.3 develops the initially designed and simplest covariance square root filter, applicable to the case of no dynamic driving noise and scalar measurements. The succeeding two sections generalize these results, first incorporating vector-valued measurements and then allowing dynamic driving noise. In Section 7.6, the square root counterpart to the inverse covariance formulation of the optimal filter is considered. Although it is not actually a square root filter, the U–D covariance factorization filter is very closely related to square root filtering, and it is depicted in Section 7.7. Finally, Section 7.8 presents the tradeoff of

numerical advantages and increased computational burden of the square root filters.

7.2 MATRIX SQUARE ROOTS

Let $\mathbf{A}$ be an n-by-n, symmetric, positive semidefinite matrix. Then there exists at least one n-by-n "square root" matrix, denoted as $\sqrt{\mathbf{A}}$, such that

$$\sqrt{\mathbf{A}}\sqrt{\mathbf{A}}^\mathsf{T} = \mathbf{A} \tag{7-1}$$

In fact, there are many matrices $\sqrt{\mathbf{A}}$ which satisfy (7-1) in general. The essential idea of square root filters is to replace the recursion for the error covariance $\mathbf{P}$ with a recursion for its square root, $\sqrt{\mathbf{P}}$, and to compute the state estimate using an optimal gain calculated in terms of $\sqrt{\mathbf{P}}$ instead of $\mathbf{P}$ itself. To motivate this, consider the scalar case: if dynamic range numerical precision problems are encountered in a filter that propagates the variance $P = \sigma^2$, the problem can be reduced by expressing the result in terms of the standard deviation σ since the dynamic range will be effectively reduced to half the original range. This basic idea can be generalized to the vector case by defining the state error covariance square roots, before and after measurement incorporation at time t_i, as $\mathbf{S}(t_i^-)$ and $\mathbf{S}(t_i^+)$ respectively, via:

$$\mathbf{S}(t_i^-)\mathbf{S}^\mathsf{T}(t_i^-) \triangleq \mathbf{P}(t_i^-) \tag{7-2}$$

$$\mathbf{S}(t_i^+)\mathbf{S}^\mathsf{T}(t_i^+) \triangleq \mathbf{P}(t_i^+) \tag{7-3}$$

Similarly, define the square root of the covariances depicting the strengths of discrete-time white Gaussian noises $\mathbf{w}_\mathrm{d}(\cdot,\cdot)$ and $\mathbf{v}(\cdot,\cdot)$ as

$$\mathbf{W}_\mathrm{d}(t_i)\mathbf{W}_\mathrm{d}^\mathsf{T}(t_i) \triangleq \mathbf{Q}_\mathrm{d}(t_i) \triangleq E\{\mathbf{w}_\mathrm{d}(t_i)\mathbf{w}_\mathrm{d}^\mathsf{T}(t_i)\} \tag{7-4}$$

$$\mathbf{V}(t_i)\mathbf{V}^\mathsf{T}(t_i) \triangleq \mathbf{R}(t_i) \triangleq E\{\mathbf{v}(t_i)\mathbf{v}^\mathsf{T}(t_i)\} \tag{7-5}$$

The covariance square roots are *not uniquely* defined by (7-2)–(7-5), and square root filters can be formulated in terms of general matrix square roots. One means of exploiting this fact is to develop algorithms which maintain a particularly attractive square root form, namely an upper or lower triangular matrix (with all zeros below or above the main diagonal, respectively), thereby requiring computation and storage of only $n(n + 1)/2$ instead of n^2 scalar variables.

This lack of uniqueness does not cause difficulties in converting from a problem description in terms of initial $\mathbf{P}_0$ and time histories of $\mathbf{Q}_\mathrm{d}(t_i)$ and $\mathbf{R}(t_i)$ to corresponding $\mathbf{S}_0$, $\mathbf{W}_\mathrm{d}(t_i)$, and $\mathbf{V}(t_i)$ values, as might first appear to be the case. The reason is that any positive semidefinite matrix can be factored into the product of a lower triangular matrix and its transpose by the *Cholesky decomposition* [13] algorithm. Although (7-1) does not uniquely define $\sqrt{\mathbf{A}}$, a *unique* Cholesky lower triangular square root $\sqrt[c]{\mathbf{A}}$ can be defined such that

$\sqrt[c]{\mathbf{A}}\sqrt[c]{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}$:

$$\begin{bmatrix} \sqrt[c]{A}_{11} & 0 & & 0 \\ \sqrt[c]{A}_{21} & \sqrt[c]{A}_{22} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt[c]{A}_{n1} & \sqrt[c]{A}_{n2} & \cdots & \sqrt[c]{A}_{nn} \end{bmatrix} \begin{bmatrix} \sqrt[c]{A}_{11} & \sqrt[c]{A}_{21} & \cdots & \sqrt[c]{A}_{n1} \\ 0 & \sqrt[c]{A}_{22} & \cdots & \sqrt[c]{A}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt[c]{A}_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{12} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}$$

The elements of the Cholesky square root matrix can be generated sequentially, row by row, from the recursion: for $i = 1, 2, \ldots, n$, compute

$$\sqrt[c]{A}_{ij} = \begin{cases} (1/\sqrt[c]{A}_{jj})[A_{ij} - \sum_{k=1}^{j-1} \sqrt[c]{A}_{ik}\sqrt[c]{A}_{jk}] & j = 1, 2, \ldots, i-1 \\ (A_{ii} - \sum_{k=1}^{i-1} \sqrt[c]{A_{ik}^2})^{1/2} & j = i \\ 0 & j > i \end{cases} \qquad (7\text{-}6)$$

Thus, $\mathbf{A}$ is scanned and $\sqrt[c]{\mathbf{A}}$ is generated in the order depicted in Fig. 7.1.

A and $\sqrt[c]{\mathbf{A}}$
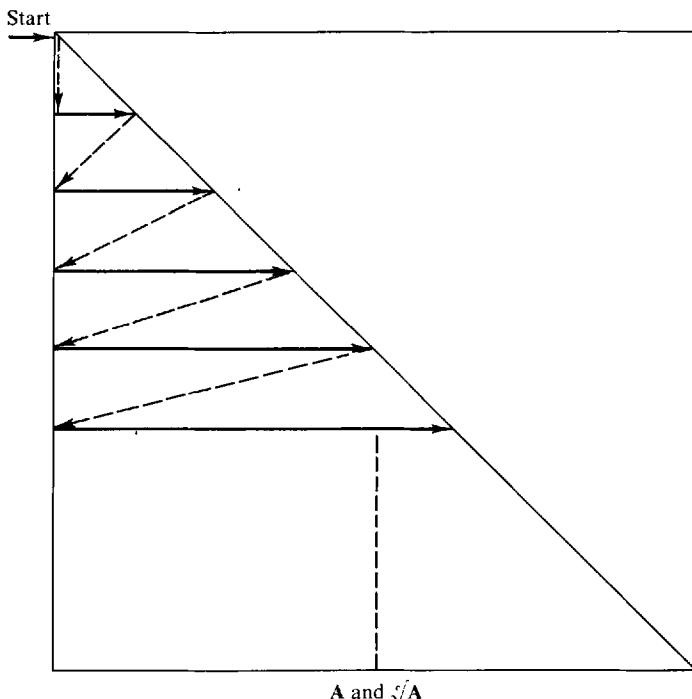
FIG. 7.1   Scanning of $\mathbf{A}$ and generation of $\sqrt[c]{\mathbf{A}}$.

EXAMPLE 7.1   Let A be given as

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 2 \\ 3 & 2 & 14 \end{bmatrix}$$

Then the elements of $\sqrt{A}$ are generated row by row as

$$\sqrt{A}_{11} = \sqrt{1} = 1, \qquad \sqrt{A}_{12} = 0, \qquad \sqrt{A}_{13} = 0$$

$$\sqrt{A}_{21} = (1/1)[2 - 0] = 2, \qquad \sqrt{A}_{22} = \sqrt{8 - 2^2} = 2, \qquad \sqrt{A}_{23} = 0$$

$$\sqrt{A}_{31} = (1/1)[3 - 0] = 3, \qquad \sqrt{A}_{32} = (1/2)[2 - (3)(2)] = -2, \qquad \sqrt{A}_{33} = [14 - (3)^2 - (-2)^2]^{1/2} = 1$$

Note that the summation term in (7-6) for $j = i$ becomes effective only for $i > 1$ and involves the sum of squares of previously generated $\sqrt{A}$ elements in that row. Furthermore, the sum term for $j < i$ is effective only for $i > 1$, and involves terms from the jth and ith rows. From above, $\sqrt{A}$ is

$$\sqrt{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & -2 & 1 \end{bmatrix}$$

and it is readily seen that $\sqrt{A}\sqrt{A}^T = A$.   ■

Later in the Carlson filter [11] of Section 7.5, we will have occasion to seek an *upper* triangular Cholesky square root $\sqrt[c]{A}$ such that $\sqrt[c]{A}\sqrt[c]{A}^T = A$. Such

A and $\sqrt[c]{A}$
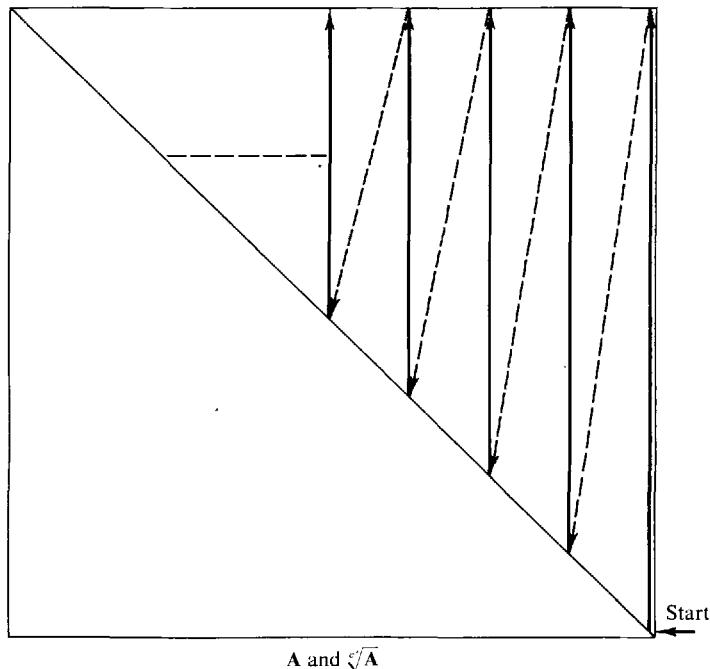
FIG. 7.2   Scanning of A and generation of $\sqrt[c]{A}$.

a matrix can be found by operating (7-6) backwards, or specifically, for $j = n$, $n - 1, \ldots, 1$, perform the following computations:

$$\sqrt[c']{A}_{ij} = \begin{cases} 0 & i > j \\ (A_{jj} - \sum_{k=j+1}^{n} \sqrt[c']{A}_{jk}^2)^{1/2} & i = j \\ (1/\sqrt[c']{A}_{jj})[A_{ij} - \sum_{k=j+1}^{n} \sqrt[c']{A}_{ik}\sqrt[c']{A}_{jk}] & i = j - 1, j - 2, \ldots, 1 \end{cases} \qquad (7\text{-}7)$$

$\sqrt[c']{A}$ is thus generated column by column, from the nth column to the first and from the bottom to top within each column, as in Fig. 7.2.

7.3   COVARIANCE SQUARE ROOT FILTER FOR $\mathbf{Q}_d \equiv 0$

In 1964, Potter [26] developed a square root filter implementation for space navigation applications in which there was no dynamic driving noise in the system model, i.e., $\mathbf{Q}_d(t_i) \equiv \mathbf{0}$ for all time, motivated by restricted wordlength in the Apollo guidance computer. For this case, the time propagation in a conventional Kalman filter would be (neglecting control inputs):

$$\hat{\mathbf{x}}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\hat{\mathbf{x}}(t_i^+) \qquad (7\text{-}8a)$$

$$\mathbf{P}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i^+)\mathbf{\Phi}^T(t_{i+1}, t_i) \qquad (7\text{-}8b)$$

By letting $\mathbf{P}(t_i^+) = \mathbf{S}(t_i^+)\mathbf{S}^T(t_i^+)$ and $\mathbf{P}(t_{i+1}^-) = \mathbf{S}(t_{i+1}^-)\mathbf{S}^T(t_{i+1}^-)$, (7-8b) can be rewritten as

$$[\mathbf{S}(t_{i+1}^-)][\mathbf{S}^T(t_{i+1}^-)] = [\mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i^+)][\mathbf{S}^T(t_i^+)\mathbf{\Phi}^T(t_{i+1}, t_i)]$$

From this it is evident that the appropriate time propagation relations for the square root filter would be

$$\hat{\mathbf{x}}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\hat{\mathbf{x}}(t_i^+) \qquad (7\text{-}9a)$$

$$\mathbf{S}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i^+) \qquad (7\text{-}9b)$$

Because of his particular application, Potter confined his attention to scalar measurements. The covariance measurement update for this case is, since $\mathbf{H}(t_i)$ is a row vector,

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{P}(t_i^-)\mathbf{H}^T(t_i) \frac{1}{[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + R(t_i)]} \mathbf{H}(t_i)\mathbf{P}(t_i^-) \quad (7\text{-}10)$$

Therefore, one can write this as

$$\mathbf{S}(t_i^+)\mathbf{S}^T(t_i^+) = \mathbf{S}(t_i^-)[\mathbf{I} - b(t_i)\mathbf{a}(t_i)\mathbf{a}^T(t_i)]\mathbf{S}^T(t_i^-) \qquad (7\text{-}11)$$

where by n-by-1 $\mathbf{a}(t_i)$ and scalar $b(t_i)$ are defined by

$$\mathbf{a}(t_i) = \mathbf{S}^T(t_i^-)\mathbf{H}^T(t_i) \qquad (7\text{-}12a)$$

$$1/b(t_i) = \mathbf{a}^T(t_i)\mathbf{a}(t_i) + R(t_i) \qquad (7\text{-}12b)$$

Potter showed that the bracketed term in (7-11) can be factored into

$$[\mathbf{I} - b\mathbf{a}\mathbf{a}^\mathrm{T}] = [\mathbf{I} - b\gamma\mathbf{a}\mathbf{a}^\mathrm{T}][\mathbf{I} - b\gamma\mathbf{a}\mathbf{a}^\mathrm{T}]^\mathrm{T} \tag{7-13}$$

where γ is a scalar defined by

$$\gamma = 1/(1 + \sqrt{bR}) \tag{7-14}$$

Substituting this into (7-11) yields the covariance update as

$$\begin{aligned}
\mathbf{S}(t_i^+) &= \mathbf{S}(t_i^-)[\mathbf{I} - b(t_i)\gamma(t_i)\mathbf{a}(t_i)\mathbf{a}^\mathrm{T}(t_i)] \\
&= \mathbf{S}(t_i^-) - b(t_i)\gamma(t_i)\mathbf{S}(t_i^-)\mathbf{a}(t_i)\mathbf{a}^\mathrm{T}(t_i)
\end{aligned} \tag{7-15}$$

The state estimate measurement update is of the conventional form, but with the Kalman gain evaluated as $[b(t_i)\mathbf{S}(t_i^-)\mathbf{a}(t_i)]$. Thus, the measurement update becomes

$$\begin{aligned}
\mathbf{a}(t_i) &= \mathbf{S}^\mathrm{T}(t_i^-)\mathbf{H}^\mathrm{T}(t_i) \\
b(t_i) &= 1/[\mathbf{a}^\mathrm{T}(t_i)\mathbf{a}(t_i) + R(t_i)] \\
\gamma(t_i) &= 1/[1 + \{b(t_i)R(t_i)\}^{1/2}] \\
\mathbf{K}(t_i) &= b(t_i)\mathbf{S}(t_i^-)\mathbf{a}(t_i) \\
\hat{\mathbf{x}}(t_i^+) &= \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \\
\mathbf{S}(t_i^+) &= \mathbf{S}(t_i^-) - \gamma(t_i)\mathbf{K}(t_i)\mathbf{a}^\mathrm{T}(t_i)
\end{aligned} \tag{7-16}$$

An equivalent form that is often employed is

$$\begin{aligned}
\mathbf{a}(t_i) &= \mathbf{S}^\mathrm{T}(t_i^-)\mathbf{H}^\mathrm{T}(t_i) \\
\sigma(t_i) &= [\mathbf{a}^\mathrm{T}(t_i)\mathbf{a}(t_i) + R(t_i)]^{1/2} \\
\alpha(t_i) &= \sigma(t_i) + V(t_i) \\
\beta(t_i) &= 1/[\sigma(t_i)\alpha(t_i)] \\
\mathbf{g}(t_i) &= \beta(t_i)[\mathbf{S}(t_i^-)\mathbf{a}(t_i)] \\
\hat{\mathbf{x}}(t_i^+) &= \hat{\mathbf{x}}(t_i^-) + \mathbf{g}(t_i)\{[\alpha(t_i)/\sigma(t_i)][z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]\} \\
\mathbf{S}(t_i^+) &= \mathbf{S}(t_i^-) - \mathbf{g}(t_i)\mathbf{a}^\mathrm{T}(t_i)
\end{aligned} \tag{7-17}$$

An example using both (7-16) and (7-17) will be presented in Section 7.5.

Note that, even if $\mathbf{S}(t_i^-)$ is lower triangular, $\mathbf{S}(t_i^+)$ will generally not be lower triangular when this update form is used. A method that preserves the lower triangular nature of the covariance square root will be discussed later.

7.4 VECTOR-VALUED MEASUREMENTS

The preceding section considered scalar measurement updates. Bellantoni and Dodge [3] extended these results to the vector measurement case by using eigenvalue decompositions, but their algorithm is inefficient for the typical

case in which the measurement vector dimension m is significantly less than the state dimension n. Andrews [2] also developed an update that processed an m-dimensional measurement vector in a single update, without requiring diagonalization:

$$\mathbf{A}(t_i) = \mathbf{S}^{\mathrm{T}}(t_i{}^-)\mathbf{H}^{\mathrm{T}}(t_i)$$
$$\mathbf{\Sigma}(t_i) = \sqrt[c]{\mathbf{A}^{\mathrm{T}}(t_i)\mathbf{A}(t_i) + \mathbf{R}(t_i)}$$
$$\hat{\mathbf{x}}(t_i{}^+) = \hat{\mathbf{x}}(t_i{}^-) + \mathbf{S}(t_i{}^-)\mathbf{A}(t_i)[\mathbf{\Sigma}^{-1}(t_i)]^{\mathrm{T}}\mathbf{\Sigma}^{-1}(t_i)[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i{}^-)] \quad (7\text{-}18)$$
$$\mathbf{S}(t_i{}^+) = \mathbf{S}(t_i{}^-) - \mathbf{S}(t_i{}^-)\mathbf{A}(t_i)[\mathbf{\Sigma}^{-1}(t_i)]^{\mathrm{T}}[\mathbf{\Sigma}(t_i) + \mathbf{V}(t_i)]^{-1}\mathbf{A}^{\mathrm{T}}(t_i)$$

This can be seen to be a direct extension of (7-17), and it is more efficient computationally than the Bellantoni and Dodge algorithm. Processing a measurement entails a Cholesky decomposition of an m-by-m matrix to generate $\mathbf{\Sigma}(t_i)$, [the extension of $\sigma(t_i)$] and inversion of two triangular m-by-m matrices, $\mathbf{\Sigma}(t_i)$ and $[\mathbf{\Sigma}(t_i) + \mathbf{V}(t_i)]$.

For the covariance square root filter, the most efficient means of performing a vector measurement update is to employ the Potter scalar update, (7-16) or (7-17), repeatedly m times. An m-dimensional measurement vector $\mathbf{z}_i$ can *always* be processed equivalently as m scalar measurements. If $\mathbf{R}(t_i)$ is diagonal, the m components can be treated as independent measurements and processed sequentially. If $\mathbf{R}(t_i)$ is not diagonal, the procedure is somewhat more complicated. First the Cholesky decomposition of $\mathbf{R}(t_i)$ is computed, yielding $\sqrt[c]{\mathbf{R}(t_i)}$ as a lower triangular matrix. Then a transformation of variables is used to convert

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \quad (7\text{-}19)$$

into

$$\mathbf{z}^*(t_i) = \mathbf{H}^*(t_i)\mathbf{x}(t_i) + \mathbf{v}^*(t_i) \quad (7\text{-}20)$$

where

$$\sqrt[c]{\mathbf{R}(t_i)}\mathbf{z}^*(t_i) = \mathbf{z}(t_i) \quad (7\text{-}21a)$$
$$\sqrt[c]{\mathbf{R}(t_i)}\mathbf{H}^*(t_i) = \mathbf{H}(t_i) \quad (7\text{-}21b)$$
$$\sqrt[c]{\mathbf{R}(t_i)}\mathbf{v}^*(t_i) = \mathbf{v}(t_i) \quad (7\text{-}21c)$$

Note that (7-21c) implies that $\mathbf{v}^*(\cdot,\cdot)$ is a unit power white Gaussian noise, i.e., $E\{\mathbf{v}^*(t_i)\mathbf{v}^{*\mathrm{T}}(t_i)\} = \mathbf{I}$, since

$$E\{\mathbf{v}(t_i)\mathbf{v}^{\mathrm{T}}(t_i)\} = \mathbf{R}(t_i) = E\{\sqrt[c]{\mathbf{R}(t_i)}\mathbf{v}^*(t_i)\mathbf{v}^{*\mathrm{T}}(t_i)\sqrt[c]{\mathbf{R}(t_i)}^{\mathrm{T}}\}$$
$$= \sqrt[c]{\mathbf{R}(t_i)}E\{\mathbf{v}^*(t_i)\mathbf{v}^{*\mathrm{T}}(t_i)\}\sqrt[c]{\mathbf{R}(t_i)}^{\mathrm{T}}$$

Thus, the components of $\mathbf{z}^*(t_i,\omega_j) = \mathbf{z}_i^*$ can be processed one at a time sequentially. Moreover, (7-21a) and (7-21b) can be solved to yield $\mathbf{z}_i^*$ and $\mathbf{H}^*(t_i)$ by simple back substitution rather than matrix inversion, because $\sqrt[c]{\mathbf{R}(t_i)}$ is

lower triangular and thus the jth component of z_i^* is a linear combination of the first j components of z_i.

EXAMPLE 7.2 Consider a four-state estimation problem with a three-dimensional vector measurement, with

$$\mathbf{H}(t_i) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \qquad \mathbf{R}(t_i) = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 2 \\ 3 & 2 & 14 \end{bmatrix}$$

Let the realized value of the measurement (7-19) be

$$\mathbf{z}(t_i, \omega_j) = \mathbf{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ z_{i3} \end{bmatrix}$$

From Example 7.1, the Cholesky square root of $\mathbf{R}(t_i)$ is

$$\sqrt{\mathbf{R}(t_i)} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & -2 & 1 \end{bmatrix}$$

The problem is equivalent to one in which a measurement of the form of (7-20) is made available, in which $\mathbf{v}^*(t_i)$ is a unit power noise. Equation (7-21a) yields

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & -2 & 1 \end{bmatrix} \begin{bmatrix} z_{i1}^* \\ z_{i2}^* \\ z_{i3}^* \end{bmatrix} = \begin{bmatrix} z_{i1} \\ z_{i2} \\ z_{i3} \end{bmatrix} \Rightarrow \begin{array}{c} z_{i1}^* = z_{i1} \\ 2z_{i1}^* + 2z_{i2}^* = z_{i2} \\ 3z_{i1}^* - 2z_{i2}^* + z_{i3}^* = z_{i3} \end{array}$$

Back substitution yields, sequentially:

$$z_{i1}^* = z_{i1}, \qquad z_{i2}^* = \tfrac{1}{2}[z_{i2} - 2z_{i1}^*], \qquad z_{i3}^* = z_{i3} - 3z_{i1}^* + 2z_{i2}^*$$

Back substitution can also be used to solve

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & -2 & 1 \end{bmatrix} \begin{bmatrix} H_{11}^* & H_{12}^* & H_{13}^* & H_{14}^* \\ H_{21}^* & H_{22}^* & H_{23}^* & H_{24}^* \\ H_{31}^* & H_{32}^* & H_{33}^* & H_{34}^* \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

row by row as

$$\begin{aligned} H_{1j}^* &= H_{1j} & j &= 1, 2, 3, 4 \\ H_{2j}^* &= \tfrac{1}{2}[H_{2j} - 2H_{1j}^*] & j &= 1, 2, 3, 4 \\ H_{3j}^* &= H_{3j} - 3H_{1j}^* + 2H_{2j}^* & j &= 1, 2, 3, 4 \end{aligned}$$

or

$$\mathbf{H}^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & \tfrac{1}{2} & 0 & 0 \\ -5 & 1 & 1 & -1 \end{bmatrix}$$

Using the transformed measurements, (7-16) or (7-17) can be applied iteratively three times to perform the update. ∎

As noted in the previous section, the $S(t_i^+)$ matrix generated by these update forms is generally not lower triangular, even if $S(t_i^-)$ is.

7.5   COVARIANCE SQUARE ROOT FILTER FOR $Q_d \neq 0$

If dynamic driving noise enters the system model, the conventional Kalman filter propagates the covariance matrix from one measurement time to the next by means of

$$P(t_{i+1}^-) = \Phi(t_{i+1}, t_i)P(t_i^+)\Phi^T(t_{i+1}, t_i) + G_d(t_i)Q_d(t_i)G_d^T(t_i) \qquad (7\text{-}22)$$

where this might be an equivalent discrete-time representation of a continuous-time system with sampled output (in which case $G_d(t_i) \equiv I$). Now we wish to develop an analogous recursion to yield $S(t_{i+1}^-)$ in terms of $S(t_i^+)$. It would be desirable to generate a lower triangular $S(t_{i+1}^-)$ since then only $\frac{1}{2}n(n + 1)$ elements would require computation rather than n^2.

One means of achieving the desired result is called the *matrix RSS (root-sum-square) method* [11]:

$$\begin{aligned} X(t_{i+1}) &= \Phi(t_{i+1}, t_i)S(t_i^+) \\ P(t_{i+1}^-) &= X(t_{i+1})X^T(t_{i+1}) + [G_d(t_i)Q_d(t_i)G_d^T(t_i)] \\ S(t_{i+1}^-) &= \sqrt[c]{P(t_{i+1}^-)} \end{aligned} \qquad (7\text{-}23)$$

This method actually computes $P(t_{i+1}^-)$ and then generates $S(t_{i+1}^-)$ as its lower triangular Cholesky square root. Although this is a rapid method, it does suffer in having only the same numerical precision as the conventional filter time propagation. Nevertheless, since it is the measurement update and not the time propagation that causes the critical numerical problems in the filter, (7-23) may well be acceptable for many applications.

EXAMPLE 7.3   Let

$$\Phi S(t_i^+) = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}, \qquad G_d Q_d G_d^T = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

Note that $[\Phi S(t_i^+)]$ has purposely been chosen as nontriangular. Equation (7-23) yields, evaluating the Cholesky square root by (7-6),

$$X(t_{i+1}) = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$$

$$P(t_{i+1}^-) = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 8 & 8 \\ 8 & 10 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix}$$

$$S_{11} = \sqrt{9} = 3, \qquad S_{12} = 0, \qquad S_{21} = \tfrac{1}{3} \cdot 9 = 3, \qquad S_{22} = \sqrt{13 - 3^2} = \sqrt{4} = 2$$

Thus

$$S(t_{i+1}^-) = \begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix} \qquad \blacksquare$$

The other means of establishing the time propagation relations is called the *triangularization method* [18]. In Section 7.3, the desired result (7-9b) was established by writing both sides of the covariance propagation (7-8b) in terms of a factor times its own transpose, and then equating the individual factors. Let us attempt to apply the same logic to (7-22). Assume that the square roots of $\mathbf{P}(t_i{}^+)$ and $\mathbf{Q}_d(t_i)$ are available: for $\mathbf{P}_0$ and $\mathbf{Q}_d(t_i)$ for all t_i, a Cholesky decomposition could be used, and for $\mathbf{P}(t_i{}^+)$ in general, assume the square root has been propagated and updated by the filter algorithm. Thus, we have

$$\mathbf{P}(t_i{}^+) = \mathbf{S}(t_i{}^+)\mathbf{S}^{\mathrm{T}}(t_i{}^+) \tag{7-24a}$$

$$\mathbf{Q}_d(t_i) = \mathbf{W}_d(t_i)\mathbf{W}_d{}^{\mathrm{T}}(t_i) \tag{7-24b}$$

Note that $\mathbf{S}(t_i{}^+)$ need not be lower triangular (important in view of the preceding section). Equation (7-22) can therefore be written as

$$\mathbf{P}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i{}^+)\mathbf{S}^{\mathrm{T}}(t_i{}^+)\mathbf{\Phi}^{\mathrm{T}}(t_{i+1}, t_i)$$
$$+ \mathbf{G}_d(t_i)\mathbf{W}_d(t_i)\mathbf{W}_d{}^{\mathrm{T}}(t_i)\mathbf{G}_d{}^{\mathrm{T}}(t_i) \tag{7-25}$$

Now it is desired to find the propagation equation for the square root of $\mathbf{P}(t_{i+1}^-)$: to find the relation to yield $\mathbf{S}(t_{i+1}^-)$ such that $\mathbf{S}(t_{i+1}^-)\mathbf{S}^{\mathrm{T}}(t_{i+1}^-)$ is equal to the right hand side of (7-25).

One such matrix would be $\tilde{\mathbf{S}}(t_{i+1}^-)$ defined by

$$\tilde{\mathbf{S}}(t_{i+1}^-) = \left[\mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i{}^+) \mid \mathbf{G}_d(t_i)\mathbf{W}_d(t_i)\right] \tag{7-26}$$

However, if $\mathbf{S}(t_i{}^+)$ is n-by-n, then $[\mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i{}^+)]$ is n-by-n and $[\mathbf{G}_d(t_i)\mathbf{W}_d(t_i)]$ is n-by-s, so $\tilde{\mathbf{S}}(t_{i+1}^-)$ would be an n-by-$(n + s)$ square root of $\mathbf{P}(t_{i+1}^-)$. Since this type of square root increases the dimension of the covariance square root matrix for each propagation interval, it must be rejected as computationally impractical.

However, this does in fact provide a fruitful insight. If $\tilde{\mathbf{S}}(t_{i+1}^-)$ is a square root of $\mathbf{P}(t_{i+1}^-)$, then so is $[\tilde{\mathbf{S}}(t_{i+1}^-)\mathbf{T}]$ if $\mathbf{T}$ is an orthogonal $(n + s)$-by-$(n + s)$ matrix, i.e., $\mathbf{T}\mathbf{T}^{\mathrm{T}} = \mathbf{I}$, since

$$\tilde{\mathbf{S}}(t_{i+1}^-)\mathbf{T}\mathbf{T}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-) = \tilde{\mathbf{S}}(t_{i+1}^-)\tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-) \tag{7-27}$$

Therefore, if an orthogonal matrix $\mathbf{T}$ can be found such that

$$\tilde{\mathbf{S}}(t_{i+1}^-)\mathbf{T} = \underbrace{\left[\mathbf{S}(t_{i+1}^-)}_{n \text{ columns}} \mid \underbrace{\mathbf{0}}_{s \text{ columns}}\right]\Big\} n \text{ rows} \tag{7-28}$$

then in fact an n-by-n square root matrix $\mathbf{S}(t_{i+1}^-)$ will have been found which satisfies the desired relationship. If, in addition, this $\mathbf{S}(t_{i+1}^-)$ were lower triangular, the result would be especially advantageous. Two methods [22] of generating such a $\mathbf{S}(t_{i+1}^-)$, known as triangularization algorithms, are the *Gram–Schmidt orthogonalization* [13, 28] procedure and the *Householder transformation* [13, 20] technique. Note that the same procedure could also

be applied to

$$\mathbf{P}(t_i{}^+) = [\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]\mathbf{P}(t_i{}^-)[\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i)]^T + \mathbf{K}(t_i)\mathbf{R}(t_i)\mathbf{K}^T(t_i)$$

or

$$\mathbf{P}^{-1}(t_i{}^+) = \mathbf{P}^{-1}(t_i{}^-) + \mathbf{H}^T(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i)$$

for vector measurement updates; the latter of these will be discussed subsequently.

First let us demonstrate that the Gram–Schmidt procedure yields the desired result. Let $\mathbf{e}^k$ denote the n-dimensional vector composed of all zeros except for a one as the kth component, so that $\mathbf{e}^1, \mathbf{e}^2, \ldots, \mathbf{e}^n$ form the standard basis for n-dimensional space. Then $[\tilde{\mathbf{S}}^T(t_{i+1}^-)\mathbf{e}^k]$ is just the kth column of $\tilde{\mathbf{S}}^T(t_{i+1}^-)$, of dimension $(n + s)$:

$$\tilde{\mathbf{S}}^T(t_{i+1}^-) = \begin{bmatrix} | & | & & | \\ \tilde{\mathbf{s}}^1 & \tilde{\mathbf{s}}^2 & \cdots & \tilde{\mathbf{s}}^n \\ | & | & & | \end{bmatrix} \Bigg\} {\scriptstyle (n+s) \text{ rows}} \qquad (7\text{-}29)$$

$$\underbrace{}_{n \text{ columns}}$$

where

$$\begin{aligned} \tilde{\mathbf{s}}^1 &= \tilde{\mathbf{S}}^T(t_{i+1}^-)\mathbf{e}^1 \\ \tilde{\mathbf{s}}^2 &= \tilde{\mathbf{S}}^T(t_{i+1}^-)\mathbf{e}^2 \\ &\vdots \\ \tilde{\mathbf{s}}^n &= \tilde{\mathbf{S}}^T(t_{i+1}^-)\mathbf{e}^n \end{aligned} \qquad (7\text{-}30)$$

Construct the orthonormal basis vectors $\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^n$ [each of dimension $(n + s)$] by the Gram–Schmidt procedure as:

$$\begin{aligned} \mathbf{b}^1 &= \text{unit } (\tilde{\mathbf{s}}^1) \\ \mathbf{b}^2 &= \text{unit } (\tilde{\mathbf{s}}^2 - [\tilde{\mathbf{s}}^{2T}\mathbf{b}^1]\mathbf{b}^1) \\ \mathbf{b}^3 &= \text{unit } (\tilde{\mathbf{s}}^3 - [\tilde{\mathbf{s}}^{3T}\mathbf{b}^1]\mathbf{b}^1 - [\tilde{\mathbf{s}}^{3T}\mathbf{b}^2]\mathbf{b}^2) \\ &\vdots \end{aligned} \qquad (7\text{-}31)$$

If $\mathbf{P}(t_i{}^+)$ is positive definite, then $\tilde{\mathbf{S}}^T(t_{i+1}^-)$ is of rank n, so n such orthogonal unit basis vectors can be obtained. Now the desired orthogonal transformation matrix $\mathbf{T}$ can be defined as the $(n + s)$-by-$(n + s)$ matrix

$$\mathbf{T} = [\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^n, \mathbf{b}^{n+1}, \ldots, \mathbf{b}^{n+s}] \qquad (7\text{-}32)$$

where $\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^n$ have been computed as in (7-31) and the remaining s columns, $\mathbf{b}^{n+1}, \ldots, \mathbf{b}^{n+s}$, are additional orthogonal unit basis vectors of $(n + s)$-dimensional space. However, it will be shown that they do not have to be computed to obtain $\mathbf{S}(t_{i+1}^-)$, so their generation will not be specified explicitly.

At this point, $\mathbf{T}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-)$ can be written as

$$
\mathbf{T}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-) =
\begin{bmatrix}
- & \mathbf{b}^{1\mathrm{T}} & - \\
- & \mathbf{b}^{2\mathrm{T}} & - \\
  & \vdots & \\
- & \mathbf{b}^{(n+s)\mathrm{T}} & -
\end{bmatrix}
\begin{bmatrix}
| & | & & | \\
\tilde{\mathbf{s}}^1 & \tilde{\mathbf{s}}^2 & \cdots & \tilde{\mathbf{s}}^n \\
| & | & & |
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
\mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^1 & \mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^2 & & \mathbf{b}^{1\mathrm{T}}\mathbf{s}^n \\
\mathbf{b}^{2\mathrm{T}}\tilde{\mathbf{s}}^1 & \mathbf{b}^{2\mathrm{T}}\tilde{\mathbf{s}}^2 & \cdots & \mathbf{b}^{2\mathrm{T}}\mathbf{s}^n \\
\vdots & \vdots & & \vdots \\
\mathbf{b}^{n\mathrm{T}}\tilde{\mathbf{s}}^1 & \mathbf{b}^{n\mathrm{T}}\tilde{\mathbf{s}}^2 & \cdots & \mathbf{b}^{n\mathrm{T}}\tilde{\mathbf{s}}^n \\
\vdots & \vdots & & \vdots \\
\mathbf{b}^{(n+s)\mathrm{T}}\tilde{\mathbf{s}}^1 & \mathbf{b}^{(n+s)\mathrm{T}}\tilde{\mathbf{s}}^2 & \cdots & \mathbf{b}^{(n+s)\mathrm{T}}\tilde{\mathbf{s}}^n
\end{bmatrix}
\tag{7-33}
$$

However, since the rank of $\tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-)$ is n and $\{\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^n\}$ span its range space while $\{\mathbf{b}^{n+1}, \ldots, \mathbf{b}^{n+s}\}$ are orthogonal to this spanning set, it follows that the last s rows in (7-33) are all zeros. By the manner in which the basis vectors were chosen, it is also true that

$$\mathbf{b}^{k\mathrm{T}}\tilde{\mathbf{s}}^j = 0, \qquad k > j$$

by the same reasoning. Thus, (7-33) becomes

$$
\mathbf{T}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-) =
\left.
\begin{bmatrix}
\mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^1 & \mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^2 & \cdots & \mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^n \\
 & \mathbf{b}^{2\mathrm{T}}\tilde{\mathbf{s}}^2 & \cdots & \mathbf{b}^{2\mathrm{T}}\tilde{\mathbf{s}}^n \\
\mathbf{0} & & \ddots & \vdots \\
 & & & \mathbf{b}^{n\mathrm{T}}\tilde{\mathbf{s}}^n \\
\hline
 & \mathbf{0} & &
\end{bmatrix}
\right\}
\begin{array}{l} n \text{ rows} \\[20pt] s \text{ rows} \end{array}
\tag{7-34}
$$

$$\underbrace{}_{n \text{ columns}}$$

The upper n-by-n partition of this matrix is just the $\mathbf{S}^{\mathrm{T}}(t_{i+1}^-)$ we have been seeking, so that $\mathbf{S}(t_{i+1}^-)$ is in fact an n-by-n lower triangular matrix:

$$
\mathbf{S}(t_{i+1}^-) =
\begin{bmatrix}
\mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^1 & & \mathbf{0} \\
\mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^2 & \mathbf{b}^{2\mathrm{T}}\tilde{\mathbf{s}}^2 & \\
\vdots & \vdots & \ddots \\
\mathbf{b}^{1\mathrm{T}}\tilde{\mathbf{s}}^n & \mathbf{b}^{2\mathrm{T}}\tilde{\mathbf{s}}^n & \cdots & \mathbf{b}^{n\mathrm{T}}\tilde{\mathbf{s}}^n
\end{bmatrix}
\tag{7-35}
$$

An efficient computational form of the Gram–Schmidt orthogonalization called the *modified Gram–Schmidt* (MGS) [22, 28] algorithm has been shown to be numerically superior to the straightforward classical procedure, i.e., less susceptible to roundoff errors [9, 27]. Moreover, it requires no more arithmetic operations than the conventional Gram–Schmidt procedure, uses less storage, and has been shown [21] to have numerical accuracy comparable to Householder [19, 23] and Givens [15] transformations. Generation of $\mathbf{S}(t_{i+1}^-)$ through this recursion proceeds as follows. Define the initial condition on $\mathbf{A}^k$, an $(n+s)$-

by-n matrix, as

$$\mathbf{A}^1 = \tilde{\mathbf{S}}^\mathrm{T}(t_{i+1}^-) = [\mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i^+) \mid \mathbf{G_d}(t_i)\mathbf{W_d}(t_i)]^T \qquad (7\text{-}36)$$

Notationally let $\mathbf{A}_j^k$ denote the jth column of $\mathbf{A}^k$. Perform the n-step recursion, for $k = 1, 2, \ldots, n$:

$$a^k = \sqrt{\mathbf{A}_k^{k\mathrm{T}}\mathbf{A}_k^{\ k}}$$

$$C_{kj} = \begin{cases} 0 & j = 1, \ldots, k-1 \\ a^k & j = k \\ [(1/a^k)\mathbf{A}_k^{k\mathrm{T}}]\mathbf{A}_j^{\ k} & j = k+1, \ldots, n \end{cases} \qquad (7\text{-}37)$$

$$\mathbf{A}_j^{k+1} = \mathbf{A}_j^{\ k} - C_{kj}[(1/a^k)\mathbf{A}_k^{\ k}] \qquad j = k+1, \ldots, n$$

Note that on successive iterations, the new $\mathbf{A}_j^{k+1}$ vectors can be "written over" the $\mathbf{A}_j^{\ k}$ vectors to conserve memory. At the end of this recursion,

$$\mathbf{S}(t_{i+1}^-) = \mathbf{C}^\mathrm{T} \qquad (7\text{-}38)$$

Notice that the computational algorithm never calculates or stores $\mathbf{T}$ explicitly in generating $\mathbf{S}(t_{i+1}^-)$.

EXAMPLE 7.4   As in Example 7.3, let

$$\mathbf{\Phi S}(t_i^+) = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}, \qquad \mathbf{G_d Q_d G_d}^\mathrm{T} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}, \qquad \mathbf{G_d W_d} = \begin{bmatrix} 1 & 0 \\ 1 & \sqrt{2} \end{bmatrix}$$

so that

$$\tilde{\mathbf{S}}(t_{i+1}^-) = [\mathbf{\Phi S}(t_i^+) \mid \mathbf{G_d W_d}] = \begin{bmatrix} 2 & 2 & 1 & 0 \\ 1 & 3 & 1 & \sqrt{2} \end{bmatrix}$$

By (7-36), the initial condition is

$$\mathbf{A}^1 = \tilde{\mathbf{S}}^T(t_{i+1}^-) = \begin{bmatrix} 2 & 1 \\ 2 & 3 \\ 1 & 1 \\ 0 & \sqrt{2} \end{bmatrix}$$

The first pass through the recursion (7-37) yields:

$$a^1 = [(2)^2 + (2)^2 + (1)^2 + (0)^2]^{1/2} = \sqrt{9} = 3$$

$$C_{11} = a^1 = 3, \qquad C_{12} = \tfrac{1}{3}[2 \cdot 1 + 2 \cdot 3 + 1 \cdot 1 + 0 \cdot \sqrt{2}] = \tfrac{1}{3}[9] = 3$$

$$\mathbf{A}_2^2 = \mathbf{A}_2^1 - C_{12}(1/a^1)\mathbf{A}_1^1 = \begin{bmatrix} 1 \\ 3 \\ 1 \\ \sqrt{2} \end{bmatrix} - (3)\tfrac{1}{3}\begin{bmatrix} 2 \\ 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ \sqrt{2} \end{bmatrix}$$

The second iteration of (7-37) produces:

$$a^2 = [(-1)^2 + (1)^2 + (0)^2 + (\sqrt{2})^2]^{1/2} = \sqrt{4} = 2$$

$$C_{21} = 0, \qquad C_{22} = 2$$

$$\mathbf{A}_3^3 = \text{not computed}$$

Finally, (7-38) generates $\mathbf{S}(t_{i+1}^-)$ as

$$\mathbf{S}(t_{i+1}^-) = \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix}$$

which agrees with the result of Example 7.3. Moreover

$$\mathbf{S}(t_{i+1}^-)\mathbf{S}^{\mathrm{T}}(t_{i+1}^-) = \begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix}\begin{bmatrix} 3 & 3 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix}$$

which agrees with a conventional Kalman filter covariance time propagation computation for this problem. ■

A *Householder transformation* [20] can also be used to solve (7-28) for the square root matrix $\mathbf{S}(t_{i+1}^-)$. Conceptually, it generates $\mathbf{T}$ as

$$\mathbf{T} = \mathbf{T}^n\mathbf{T}^{(n-1)}\dots\mathbf{T}^1$$

where $\mathbf{T}^k$ is generated recursively as

$$\mathbf{T}^k = \mathbf{I} - d^k\mathbf{u}^k\mathbf{u}^{k\mathrm{T}}$$

with the scalar d^k and the $(n+s)$-vector $\mathbf{u}^k$ defined in the following. However, the computational algorithm never calculates these $\mathbf{T}^k$'s or $\mathbf{T}$ explicitly. The initial condition on the $(n+s)$-by-n $\mathbf{A}^k$ is

$$\mathbf{A}^1 = \tilde{\mathbf{S}}^{\mathrm{T}}(t_{i+1}^-) = [\boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i^+) \mid \mathbf{G}_{\mathrm{d}}(t_i)\mathbf{W}_{\mathrm{d}}(t_i)]^{\mathrm{T}} \qquad (7\text{-}39)$$

Again, letting $\mathbf{A}_j^{\ k}$ represent the jth column of $\mathbf{A}^k$, perform the n-step recursion, for $k = 1, 2, \dots, n$:

$$a^k = \sqrt{\sum_{j=k}^{n+s}[A_{jk}^k]^2}\cdot\mathrm{sgn}\{A_{kk}^k\}$$

$$d^k = \frac{1}{a^k(a^k + A_{kk}^k)}$$

$$u_j^{\ k} = \begin{cases} 0 & j < k \\ a^k + A_{kk}^k & j = k \\ A_{jk}^k & j = (k+1), \dots, (n+s) \end{cases}$$

$$y_j^{\ k} = \begin{cases} 0 & j < k \\ 1 & j = k \\ d^k\mathbf{u}^{k\mathrm{T}}\mathbf{A}_j^{\ k} & j = (k+1), \dots, n \end{cases}$$ $\qquad (7\text{-}40)$

$$\mathbf{A}^{k+1} = \mathbf{A}^k - \mathbf{u}^k\mathbf{y}^{k\mathrm{T}}$$

At stage k, the first $(k-1)$ columns of $\mathbf{A}^k$ are zero below the diagonal of the upper square partition, and $\mathbf{u}^k$ has been chosen so that the subdiagonal elements of $\mathbf{A}_k^{k+1}$ will be zero. After the n iterations of (7-40),

$$\mathbf{A}^{n+1} = \begin{bmatrix} \mathbf{C} \\ \hline \mathbf{0} \end{bmatrix}\begin{matrix} \}n \\ \}s \end{matrix} \qquad (7\text{-}41)$$

and then $\mathbf{S}(t_{i+1}^-)$ is generated as

$$\mathbf{S}(t_{i+1}^-) = \mathbf{C}^{\mathrm{T}} \tag{7-42}$$

EXAMPLE 7.5   Consider the same problem as in Example 7.4. By (7-39),

$$\mathbf{A}^1 = \begin{bmatrix} 2 & 1 \\ 2 & 3 \\ 1 & 1 \\ 0 & \sqrt{2} \end{bmatrix}$$

The first iteration of (7-40) yields

$$a^1 = (2^2 + 2^2 + 1^2 + 0^2)^{1/2} \cdot \mathrm{sgn}\{2\} = \sqrt{9} = 3, \qquad d^1 = 1/[3(3+2)] = 1\ 15$$

$$u_1{}^1 = (3+2) = 5, \qquad u_2{}^1 = 2, \qquad u_3{}^1 = 1, \qquad u_4{}^1 = 0$$

$$y_1{}^1 = 1, \qquad y_2{}^1 = (1/15)[5 \cdot 1 + 2 \cdot 3 + 1 \cdot 1 + 0 \cdot \sqrt{2}] = 4/5$$

$$\mathbf{A}^2 = \begin{bmatrix} 2 & 1 \\ 2 & 3 \\ 1 & 1 \\ 0 & \sqrt{2} \end{bmatrix} - \begin{bmatrix} 5 \\ 2 \\ 1 \\ 0 \end{bmatrix}[1 \quad 4/5] = \begin{bmatrix} -3 & -3 \\ 0 & 7/5 \\ 0 & 1/5 \\ 0 & \sqrt{2} \end{bmatrix}$$

The second iteration of (7-40) produces

$$a^2 = [(7/5)^2 + (1/5)^2 + \sqrt{2}^2]^{1/2} \cdot \mathrm{sgn}\{7/5\} = \sqrt{4} = 2, \qquad d^2 = 1/\{2[2 + (7/5)]\} = 5/34$$

$$u_1{}^2 = 0, \qquad u_2{}^2 = 2 + (7/5) = 17/5, \qquad u_3{}^2 = 1/5, \qquad u_4{}^2 = \sqrt{2}$$

$$y_1{}^2 = 0, \qquad y_2{}^2 = 1$$

$$\mathbf{A}^3 = \begin{bmatrix} -3 & -3 \\ 0 & 7/5 \\ 0 & 1/5 \\ 0 & \sqrt{2} \end{bmatrix} - \begin{bmatrix} 0 \\ 17/5 \\ 1/5 \\ \sqrt{2} \end{bmatrix}[0 \quad 1] = \begin{bmatrix} -3 & -3 \\ 0 & -2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Therefore, from (7-41) and (7-42), $\mathbf{S}(t_{i+1}^-)$ is identified as the transpose of the upper 2-by-2 partition of $\mathbf{A}^3$:

$$\mathbf{S}(t_{i+1}^-) = \begin{bmatrix} -3 & 0 \\ -3 & -2 \end{bmatrix}$$

This is just the negative of the previous results, and thus is also a valid covariance square root. ■

The Householder triangularization requires $[4n^3 + 6n^2(s+1) + 2n]/6$ multiplies, $[4n^3 + 6sn^2 + 8n]/6$ adds, n divides, and n square roots. This is $[2n^3 - 8n]/6$ fewer multiples and $[2n^3 + 3n^2 - 11n]/6$ fewer adds than required by the modified Gram–Schmidt algorithm. However, the MGS algorithm becomes slightly more precise numerically as the residual size increases [22], and thus is a viable alternative.

A Householder transformation method has also been proposed for performing measurement updates [12, 22]. However, this has been shown to be

equivalent to the Potter method described previously, but not as efficient computationally [4].

Thus, the *covariance square root filter* (*Potter filter*) algorithm can be specified as follows. The propagation of the state estimate from one sample time to the next is given by (7-9a). Covariance square root time propagations are calculated by means of the matrix RSS method (7-23), the MGS algorithm given by (7-36)–(7-38), or the Householder transformation as in (7-39)–(7-42). Of these, the latter two are preferable since they are more accurate numerically than the computationally efficient first method, and numerics are the basic motivation for square root forms. Measurement updates would be processed through m iterations of the Potter algorithm (7-16) or (7-17). If the $\mathbf{R}(t_i)$ matrix is not diagonal, the transformation of variables given by (7-19)–(7-21) must first be performed.

EXAMPLE 7.6   This example illustrates one complete recursion of the Potter filter. Let $\mathbf{S}(t_i^-)$ and the corresponding $\mathbf{P}(t_i^-) = \mathbf{S}(t_i^-)\mathbf{S}^T(t_i^-)$ be

$$\mathbf{S}(t_i^-) = \begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix}, \qquad \mathbf{P}(t_i^-) = \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix}$$

as computed by the time propagation of Examples 7.3, 7.4, or 7.5. Now let a scalar measurement be taken such that $\mathbf{H}(t_i) = \begin{bmatrix} 1/3 & 1 \end{bmatrix}$ and $R(t_i) = 4$. A conventional Kalman update would yield

$$\mathbf{K}(t_i) = \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + R(t_i)]^{-1}$$

$$= \frac{1}{20+4}\begin{bmatrix} 12 \\ 16 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 2/3 \end{bmatrix}$$

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i^-)$$

$$= \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix} - \begin{bmatrix} 1/2 \\ 2/3 \end{bmatrix}\begin{bmatrix} 12 & 16 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 7/3 \end{bmatrix}$$

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$$

The corresponding result given by (7-16) is:

$$\mathbf{a}(t_i) = \begin{bmatrix} 3 & 3 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1/3 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

$$b(t_i) = 1/[4^2 + 2^2 + 4] = 1/24$$

$$\gamma(t_i) = 1/[1 + \sqrt{(1/24)(4)}] = 1/[1 + \sqrt{1/6}]$$

$$\mathbf{K}(t_i) = \frac{1}{24}\begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 12/24 \\ 16/24 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 2/3 \end{bmatrix}$$

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$$

$$\mathbf{S}(t_i^+) = \begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix} - \frac{1}{1 + \sqrt{1/6}}\begin{bmatrix} 1/2 \\ 2/3 \end{bmatrix}\begin{bmatrix} 4 & 2 \end{bmatrix}$$

$$= \frac{1}{1 + \sqrt{1/6}}\begin{bmatrix} 1 + 3\sqrt{1/6} & -1 \\ (1/3) + 3\sqrt{1/6} & (2/3) + 2\sqrt{1/6} \end{bmatrix}$$

Note that the computed gains $\mathbf{K}(t_i)$ agree and that

$$\mathbf{S}(t_i{}^+)\mathbf{S}^\mathrm{T}(t_i{}^+) = \begin{bmatrix} 3 & 1 \\ 1 & 7/3 \end{bmatrix}$$

which agrees with $\mathbf{P}(t_i{}^+)$.

By comparison, (7-17) yields:

$$\mathbf{a}(t_i) = \begin{bmatrix} 3 & 3 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1/3 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

$$\sigma(t_i) = (4^2 + 2^2 + 4)^{1/2} = \sqrt{24} = 2\sqrt{6}$$

$$\alpha(t_i) = 2\sqrt{6} + 2$$

$$\beta(t_i) = 1/[(2\sqrt{6})(2\sqrt{6} + 2)] = 1/[24 + 4\sqrt{6}]$$

$$\mathbf{g}(t_i) = \frac{1}{[24 + 4\sqrt{6}]}\begin{bmatrix} 12 \\ 16 \end{bmatrix} = \frac{1}{6 + \sqrt{6}}\begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$\hat{\mathbf{x}}(t_i{}^+) = \hat{\mathbf{x}}(t_i{}^-) + \frac{1}{6 + \sqrt{6}}\begin{bmatrix} 3 \\ 4 \end{bmatrix}\frac{2\sqrt{6} + 2}{2\sqrt{6}}[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i{}^-)]$$

$$\mathbf{S}(t_i{}^+) = \begin{bmatrix} 3 & 0 \\ 3 & 2 \end{bmatrix} - \frac{1}{6 + \sqrt{6}}\begin{bmatrix} 3 \\ 4 \end{bmatrix}[4 \quad 2]$$

$$= \frac{1}{1 + \sqrt{1/6}}\begin{bmatrix} 1 + 3\sqrt{1/6} & -1 \\ (1/3) + 3\sqrt{1/6} & (2/3) + 2\sqrt{1/6} \end{bmatrix}$$

The $\mathbf{S}(t_i{}^+)$ agrees with that just obtained. Moreover, if $\mathbf{g}(t_i)[\alpha(t_i)/\sigma(t_i)]$ were computed instead of the more efficient multiplication of the residual by the scalar $[\alpha(t_i)/\sigma(t_i)]$ followed by multiplication by $\mathbf{g}(t_i)$, the result would be identical to the $\mathbf{K}(t_i)$ previously computed.   ∎

One significant drawback of the covariance square root filter just described is that the triangularity of the square root matrix is generally destroyed during the measurement updating. Consequently, all n^2 elements must be computed and stored. A more recent algorithm, the *Carlson filter* [11], provides substantial improvement in both computational speed and required storage by maintaining the covariance square root matrix in triangular form. By doing so, only $n(n + 1)/2$ memory locations need be allocated for $\mathbf{S}(t_i{}^+)$, and the product $[\mathbf{\Phi}(t_{i+1}, t_i)\mathbf{S}(t_i{}^+)]$ for the subsequent time propagation requires only half the usual number of computations.

Like the Potter measurement update, the Carlson algorithm processes vector measurements iteratively as scalars. Therefore, consider the general square root solution to (7-11):

$$\mathbf{S}(t_i{}^+) = \mathbf{S}(t_i{}^-)[\mathbf{I} - b(t_i)\mathbf{a}(t_i)\mathbf{a}^\mathrm{T}(t_i)]^{1/2}$$
$$= \mathbf{S}(t_i{}^-)[\mathbf{I} - \mathbf{a}(t_i)\mathbf{a}^\mathrm{T}(t_i)/d(t_i)]^{1/2} \qquad (7\text{-}43)$$

where, for convenience, $d(t_i)$ has been defined as $1/b(t_i)$. Assuming $\mathbf{S}(t_i{}^-)$ to be upper triangular, we seek a matrix $[\mathbf{I} - \mathbf{a}(t_i)\mathbf{a}^\mathrm{T}(t_i)/d(t_i)]^{1/2}$ such that the $\mathbf{S}(t_i{}^+)$ computed in (7-43) is also upper triangular. The choice between upper and

lower triangular form is arbitrary, governed by selecting either forward or backward recursion algorithms for the Cholesky, Householder, and Gram–Schmidt procedures. Upper triangular forms are motivated to some extent by state vector partitioning, discussed in Problem 7.13.

The desired square root matrix is in fact derived by means of an analytic Cholesky decomposition, and can be expressed as

$$\begin{bmatrix} b_1 & & & \\ & b_2 & & \\ & & \ddots & \\ & & & b_n \end{bmatrix} - \begin{bmatrix} 0 & a_1 & a_1 & \cdots & a_1 \\ & 0 & a_2 & \cdots & a_2 \\ & & 0 & & \vdots \\ & & & \ddots & a_{n-1} \\ & & & & 0 \end{bmatrix} \begin{bmatrix} c_1 & & & \\ & c_2 & & \\ & & \ddots & \\ & & & c_n \end{bmatrix}$$

where a_k is the kth component of $\mathbf{a}(t_i)$, and b_k and c_k for $k = 1, 2, \ldots, n$ will be described presently. However, the computational algorithm neither computes this square root explicitly nor requires a matrix multiplication as in (7-43) to generate $\mathbf{S}(t_i^+)$. The algorithm is initialized by setting the scalar d_0 and n-vectors $\mathbf{e}_0$ and $\mathbf{a}$ as

$$d_0 = R(t_i), \qquad \mathbf{e}_0 = \mathbf{0}, \qquad \mathbf{a} = \mathbf{S}^T(t_i^-)\mathbf{H}^T(t_i) \tag{7-44}$$

and iterating for $k = 1, 2, \ldots, n$ on

$$\begin{aligned} d_k &= d_{k-1} + a_k{}^2 \\ b_k &= (d_{k-1}/d_k)^{1/2} \\ c_k &= a_k/(d_{k-1}d_k)^{1/2} \\ \mathbf{e}_k &\doteq \mathbf{e}_{k-1} + \mathbf{S}_k{}^- a_k \\ \mathbf{S}_k{}^+ &= \mathbf{S}_k{}^- b_k - \mathbf{e}_{k-1} c_k \end{aligned} \tag{7-45}$$

In the recursion, $\mathbf{S}_k{}^-$ denotes the kth column of $\mathbf{S}(t_i^-)$, and both it and $\mathbf{e}_k$ consist of zeros below the kth element. After the n iterations, $\mathbf{S}(t_i^+)$ is produced as

$$\mathbf{S}(t_i^+) = [\mathbf{S}_1{}^+ \, \mathbf{S}_2{}^+ \cdots \mathbf{S}_n{}^+] \tag{7-46}$$

which is an upper triangular matrix. The state vector update is then given by

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{e}_n\{[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]/d_n\} \tag{7-47}$$

EXAMPLE 7.7   Consider the same problem as in Example 7.6, but now assume $\mathbf{S}(t_i^-)$ to be upper triangular and such that $\mathbf{S}(t_i^-)\mathbf{S}^T(t_i^-)$ is equal to

$$\mathbf{P}(t_i^-) = \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix}$$

The upper triangular Cholesky square root is found through (7-7) as:

$$\mathbf{S}(t_i^-) = \begin{bmatrix} 6/\sqrt{13} & 9/\sqrt{13} \\ 0 & \sqrt{13} \end{bmatrix}$$

The initialization of (7-44) yields

$$d_0 = 4, \qquad \mathbf{e}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \mathbf{a} = \begin{bmatrix} 6/\sqrt{13} & 0 \\ 9/\sqrt{13} & \sqrt{13} \end{bmatrix} \begin{bmatrix} 1/3 \\ 1 \end{bmatrix} = \begin{bmatrix} 2/\sqrt{13} \\ 16/\sqrt{13} \end{bmatrix}$$

The first recursion of (7-45) yields

$$d_1 = 4 + \frac{4}{13} = \frac{56}{13}, \qquad b_1 = \left(4 \Big/ \frac{56}{13} \right)^{1/2} = \sqrt{\frac{13}{14}}$$

$$c_1 = \left[\frac{2}{\sqrt{13}} \right] \Big/ \left[4 \cdot \frac{56}{13} \right]^{1/2} = \frac{1}{2\sqrt{14}}$$

$$\mathbf{e}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 6/\sqrt{13} \\ 0 \end{bmatrix} \frac{2}{\sqrt{13}} = \begin{bmatrix} 12/13 \\ 0 \end{bmatrix}$$

$$\mathbf{S}_1^+ = \begin{bmatrix} 6/\sqrt{13} \\ 0 \end{bmatrix} \sqrt{\frac{13}{14}} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 6/\sqrt{14} \\ 0 \end{bmatrix}$$

The second iteration yields

$$d_2 = \frac{56}{13} + \frac{256}{13} = \frac{312}{13}, \qquad b_2 = \left(\frac{56}{13} \Big/ \frac{312}{13} \right)^{1/2} = \sqrt{\frac{56}{312}} = \sqrt{\frac{7}{39}}$$

$$c_2 = \left[\frac{16}{\sqrt{13}} \right] \Big/ \left[\frac{56}{13} \cdot \frac{312}{13} \right]^{1/2} = \frac{2}{\sqrt{21}}$$

$$\mathbf{e}_2 = \begin{bmatrix} 12/13 \\ 0 \end{bmatrix} + \begin{bmatrix} 9/\sqrt{13} \\ \sqrt{13} \end{bmatrix} \frac{16}{\sqrt{13}} = \begin{bmatrix} 156/13 \\ 16 \end{bmatrix}$$

$$\mathbf{S}_2^+ = \begin{bmatrix} 9/\sqrt{13} \\ \sqrt{13} \end{bmatrix} \frac{\sqrt{7}}{\sqrt{39}} - \begin{bmatrix} 12/13 \\ 0 \end{bmatrix} \frac{2}{\sqrt{21}} = \begin{bmatrix} \sqrt{3/7} \\ \sqrt{7/3} \end{bmatrix}$$

Thus, $\mathbf{S}(t_i^+)$ and $\hat{\mathbf{x}}(t_i^+)$ are given by (7-46) and (7-47) as

$$\mathbf{S}(t_i^+) = \begin{bmatrix} 6/\sqrt{14} & \sqrt{3/7} \\ 0 & \sqrt{7/3} \end{bmatrix}$$

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \begin{bmatrix} 156/13 \\ 16 \end{bmatrix} \left\{ [z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)] \Big/ \frac{312}{13} \right\}$$

Note that the value of $[\mathbf{e}_2/d_2]$, not calculated explicitly above, agrees with $\mathbf{K}(t_i)$ of Example 7.6. Moreover, $\mathbf{S}(t_i^+)\mathbf{S}^T(t_i^+)$ is equal to the $\mathbf{P}(t_i^+)$ in that example. ∎

For time propagations, Carlson suggested the matrix RSS method, (7-23), but with an upper triangular Cholesky square root as generated in (7-7) replacing the lower triangular form in (7-23). However, the triangularization methods could also be employed, thereby sacrificing some computational speed for increased numerical precision. A modified Gram–Schmidt algorithm [29] can

be written by initializing $\mathbf{A}$ according to (7-36) and then iterating for $k = n$, $n - 1, \ldots, 1$ on

$$S_{kk}(t_{i+1}^-) = \sqrt{\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k}$$
$$\mathbf{v}_k = \mathbf{A}_k/S_{kk}(t_{i+1}^-)$$
$$\left.\begin{array}{l} S_{jk}(t_{i+1}^-) = \mathbf{A}_j^{\mathsf{T}}\mathbf{v}_k \\ \mathbf{A}_j \leftarrow \mathbf{A}_j - S_{jk}(t_{i+1}^-)\mathbf{v}_k \end{array}\right\} \quad j = 1, 2, \ldots, (k-1)$$

$$(7\text{-}48)$$

where $\leftarrow$ denotes replacement by means of "writing over" old variables.

7.6 INVERSE COVARIANCE SQUARE ROOT FILTER

In Section 5.7, the inverse covariance formulation of the optimal filter was presented, an algorithm which is algebraically equivalent to the Kalman filter but has substantially different characteristics, such as being able to incorporate unknown initial conditions and being more efficient if the measurement vector is very large in dimension $(m > n)$. Now we consider the square root filter analog of such a formulation.

If we define the covariance square root matrix $\mathbf{S}(t_i^+)$ through

$$\mathbf{P}(t_i^+) \triangleq \mathbf{S}(t_i^+)\mathbf{S}^{\mathsf{T}}(t_i^+) \tag{7-49}$$

then it is consistent that an inverse covariance square root $\mathbf{S}^{-1}(t_i^+)$ be defined through

$$\mathbf{P}^{-1}(t_i^+) \triangleq \mathbf{S}^{-\mathsf{T}}(t_i^+)\mathbf{S}^{-1}(t_i^+) \tag{7-50a}$$

where $\mathbf{S}^{-\mathsf{T}}$ denotes $[\mathbf{S}^{-1}]^{\mathsf{T}} = [\mathbf{S}^{\mathsf{T}}]^{-1}$. Similarly, $\mathbf{S}^{-1}(t_i^-)$ would be defined through

$$\mathbf{P}^{-1}(t_i^-) \triangleq \mathbf{S}^{-\mathsf{T}}(t_i^-)\mathbf{S}^{-1}(t_i^-) \tag{7-50b}$$

To develop the inverse covariance square root filter [5, 22], first consider the measurement update equation in the inverse covariance filter:

$$\mathbf{P}^{-1}(t_i^+) = \mathbf{P}^{-1}(t_i^-) + \mathbf{H}^{\mathsf{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{H}(t_i) \tag{7-51}$$

Using (7-5) and (7-50), this can be written as

$$\mathbf{P}^{-1}(t_i^+) = \mathbf{S}^{-\mathsf{T}}(t_i^-)\mathbf{S}^{-1}(t_i^-) + \mathbf{H}^{\mathsf{T}}(t_i)\mathbf{V}^{-\mathsf{T}}(t_i)\mathbf{V}^{-1}(t_i)\mathbf{H}(t_i) \tag{7-52}$$

We now seek an update relation for $\mathbf{S}^{-1}(t_i^+)$ such that $\{[\mathbf{S}^{-1}(t_i^+)]^{\mathsf{T}}[\mathbf{S}^{-1}(t_i^+)]\}$ is equivalent to the right hand side of (7-52). One such matrix would be the $(n + m)$-by-n matrix

$$\tilde{\mathbf{S}}^{-1}(t_i^+) = \left[\begin{array}{c} \mathbf{S}^{-1}(t_i^-) \\ \hline \mathbf{V}^{-1}(t_i)\mathbf{H}(t_i) \end{array}\right] \tag{7-53}$$

As in the previous section, such an $\tilde{\mathbf{S}}^{-1}(t_i{}^+)$ would be unacceptable due to the increasing matrix dimensions it would cause. However, if an orthogonal matrix $\mathbf{T}$ can be constructed such that

$$\mathbf{T}^{\mathrm{T}}\tilde{\mathbf{S}}^{-1}(t_i{}^+) = \left[\begin{array}{c} \mathbf{S}^{-1}(t_i{}^+) \\ \hline \mathbf{0} \end{array}\right] \begin{array}{l} \} \ n \ \text{rows} \\ \} \ m \ \text{rows} \end{array} \tag{7-54}$$

or

$$\tilde{\mathbf{S}}^{-\mathrm{T}}(t_i{}^+)\mathbf{T} = [\mathbf{S}^{-\mathrm{T}}(t_i{}^+) \mid \mathbf{0}] \tag{7-54'}$$

then the resulting n-by-n $\mathbf{S}^{-1}(t_i{}^+)$ is the desired square root matrix. In analogy to the previous development, it would be especially beneficial if the $\mathbf{S}^{-1}(t_i{}^+)$ so generated were upper triangular. Either the modified Gram–Schmidt orthogonalization procedure or the Householder transformation algorithm can be employed to solve for the desired $\mathbf{S}^{-1}(t_i{}^+)$, and this will be developed in detail after the state estimate is discussed.

Recall from Section 5.7 that the inverse covariance filter did not compute a state estimate directly, but rather

$$\hat{\mathbf{y}}(t_i{}^-) \triangleq \mathbf{P}^{-1}(t_i{}^-)\hat{\mathbf{x}}(t_i{}^-) \tag{7-55a}$$

$$\hat{\mathbf{y}}(t_i{}^+) \triangleq \mathbf{P}^{-1}(t_i{}^+)\hat{\mathbf{x}}(t_i{}^+) \tag{7-55b}$$

which were related by

$$\hat{\mathbf{y}}(t_i{}^+) = \hat{\mathbf{y}}(t_i{}^-) + \mathbf{H}^{\mathrm{T}}(t_i)\mathbf{R}^{-1}(t_i)\mathbf{z}_i \tag{7-56}$$

Analogously, the inverse covariance square root filter does not generate an estimate of the state explicitly, but instead calculates

$$\hat{\boldsymbol{\alpha}}(t_i{}^-) \triangleq \mathbf{S}^{-1}(t_i{}^-)\hat{\mathbf{x}}(t_i{}^-) \tag{7-57a}$$

and

$$\hat{\boldsymbol{\alpha}}(t_i{}^+) \triangleq \mathbf{S}^{-1}(t_i{}^+)\hat{\mathbf{x}}(t_i{}^+) \tag{7-57b}$$

The update relationship between these estimates can be shown to be

$$\begin{array}{l} n \ \text{rows} \ \{ \\ m \ \text{rows} \ \{ \end{array} \left[\begin{array}{c} \hat{\boldsymbol{\alpha}}(t_i{}^+) \\ \hline \boldsymbol{\beta}(t_i) \end{array}\right] = \mathbf{T}^{\mathrm{T}} \left[\begin{array}{c} \hat{\boldsymbol{\alpha}}(t_i{}^-) \\ \hline \mathbf{V}^{-1}(t_i)\mathbf{z}_i \end{array}\right] \begin{array}{l} \} \ n \ \text{rows} \\ \} \ m \ \text{rows} \end{array} \tag{7-58}$$

where $\mathbf{T}$ is the same orthogonal matrix as in (7-54), and $\boldsymbol{\beta}(t_i)$ is an m-dimensional vector (the residual after processing the measurement, $[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i{}^+)]$) that need not be calculated. Since the first n rows of $\mathbf{T}^{\mathrm{T}}$ are the result of an n-step Gram–Schmidt or Householder process, $\hat{\boldsymbol{\alpha}}(t_i{}^+)$ can be computed without knowledge of any additional portion of $\mathbf{T}^{\mathrm{T}}$ than that generated by either of the triangularization algorithms discussed previously.

The *modified Gram–Schmidt* (MGS) *measurement update* initializes an $(n + m)$-by-n matrix $\mathbf{A}^k$ and an n-vector $\mathbf{b}^k$ as

$$\mathbf{A}^1 = \tilde{\mathbf{S}}^{-1}(t_i{}^+) = \left[\begin{array}{c} \mathbf{S}^{-1}(t_i^-) \\ \hline \mathbf{V}^{-1}(t_i)\mathbf{H}(t_i) \end{array} \right] \tag{7-59a}$$

$$\mathbf{b}^1 = \mathbf{0} \tag{7-59b}$$

Then an n-step recursion is performed that is identical to (7-37) except for two additional equations for eventual generation of $\hat{\mathbf{x}}(t_i{}^+)$; for $k = 1, 2, \ldots, n$,

$$a^k = \sqrt{\mathbf{A}_k^{k\mathrm{T}}\mathbf{A}_k{}^k}$$

$$C_{kj} = \begin{cases} 0 & j = 1, \ldots, k - 1 \\ a^k & j = k \\ [(1/a^k)\mathbf{A}_k^{k\mathrm{T}}]\mathbf{A}_j{}^k & j = k + 1, \ldots, n \end{cases}$$

$$e^k = [(1/a^k)\mathbf{A}_k^{k\mathrm{T}}]\mathbf{b}^k \tag{7-60}$$

$$\mathbf{A}_j^{k+1} = \mathbf{A}_j{}^k - C_{kj}[(1/a^k)\mathbf{A}_k{}^k] \quad j = k + 1, \ldots, n$$

$$\mathbf{b}^{k+1} = \mathbf{b}^k - e^k[(1/a^k)\mathbf{A}_k{}^k]$$

At the end of this recursion,

$$\mathbf{S}^{-1}(t_i{}^+) = \mathbf{C}, \qquad \hat{\mathbf{x}}(t_i{}^+) = \mathbf{b}^{n+1} \tag{7-61}$$

A *Householder measurement update* [10, 18] can also be employed. The $(n + m)$-by-n matrix $\mathbf{A}^k$ and $(n + m)$-vector $\mathbf{b}^k$ (note the different dimension on $\mathbf{b}^k$) are initialized as in (7-59). Subsequently an n-step recursion identical in form to (7-40) except for auxiliary steps to calculate $\hat{\mathbf{x}}(t_i{}^+)$ is performed [22]; for $k = 1, 2, \ldots, n$,

$$a^k = \sqrt{\sum_{j=k}^{n+m} [A_{jk}^k]^2} \cdot \mathrm{sgn}\{A_{kk}^k\}$$

$$d^k = \frac{1}{a^k(a^k + A_{kk}^k)}$$

$$u_j{}^k = \begin{cases} 0 & j < k \\ a^k + A_{kk}^k & j = k \\ A_{jk}^k & j = k + 1, \ldots, (n + m) \end{cases}$$

$$y_j{}^k = \begin{cases} 0 & j < k \\ 1 & j = k \\ d^k \mathbf{u}^{k\mathrm{T}}\mathbf{A}_j{}^k & j = k + 1, \ldots, n \end{cases} \tag{7-62}$$

$$e^k = a^k \mathbf{u}^{k\mathrm{T}}\mathbf{b}^k$$

$$\mathbf{A}^{k+1} = \mathbf{A}^k - \mathbf{u}^k \mathbf{y}^{k\mathrm{T}}$$

$$\mathbf{b}^{k+1} = \mathbf{b}^k - \mathbf{u}^k e^k$$

After the n iterations of (7-62), $S^{-1}(t_i^+)$ and $\hat{\alpha}(t_i^+)$ are obtained from

$$\mathbf{A}^{n+1} = \begin{bmatrix} \mathbf{S}^{-1}(t_i^+) \\ \hline \mathbf{0} \end{bmatrix}, \qquad \mathbf{b}^{n+1} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}(t_i^+) \\ \hline \boldsymbol{\beta}(t_i) \end{bmatrix} \tag{7-63}$$

For *time propagations* in which the dynamic driving noise is s dimensional, s scalar recursions analogous to the Potter measurement update in the covariance square root filter are performed. Thus $\mathbf{Q}_d(t_i)$ is assumed diagonal, perhaps after a change of variables (as explicitly described in the next section), and the effects of $\mathbf{w}_d(\cdot, \cdot)$ are incorporated component by component. Letting $\mathbf{G}_{dk}$ be the kth column of $\mathbf{G}_d(t_i)$ and Q_{dk} be the kth diagonal element of $\mathbf{Q}_d(t_i)$, the algorithm becomes, for $k = 1, 2, \ldots, s$,

$$\begin{aligned}
\mathbf{a}(t_i) &= \mathbf{S}^{-1}(t_i^+)\boldsymbol{\Phi}(t_i, t_{i+1})\mathbf{G}_{dk} \\
b(t_i) &= 1/[\mathbf{a}^T(t_i)\mathbf{a}(t_i) + \{1/Q_{dk}\}] \\
\gamma(t_i) &= 1/[1 + \sqrt{b(t_i)\{1/Q_{dk}\}}] \\
\mathbf{l}^T(t_i) &= b(t_i)\mathbf{a}^T(t_i)\mathbf{S}^{-1}(t_i^+)\boldsymbol{\Phi}(t_i, t_{i+1}) \\
\hat{\boldsymbol{\alpha}}(t_{i+1}^-) &= \boldsymbol{\alpha}(t_i^+) - b(t_i)\gamma(t_i)\mathbf{a}(t_i)\mathbf{a}^T(t_i)\boldsymbol{\alpha}(t_i^+) \\
\mathbf{S}^{-1}(t_{i+1}^-) &= \mathbf{S}^{-1}(t_i^+)\boldsymbol{\Phi}(t_i, t_{i+1}) - \gamma(t_i)\mathbf{a}(t_i)\mathbf{l}^T(t_i)
\end{aligned} \tag{7-64}$$

Note the order of the time indices on the state transition matrix and that $\boldsymbol{\Phi}(t_i, t_{i+1}) = \boldsymbol{\Phi}^{-1}(t_{i+1}, t_i)$. After the first of the s iterations of (7-64), $\boldsymbol{\Phi}(t_i, t_{i+1})$ is replaced by the identity matrix and $\mathbf{S}^{-1}(t_i^+)$ and $\hat{\boldsymbol{\alpha}}(t_i^+)$ are replaced by the $\mathbf{S}^{-1}$ and $\hat{\boldsymbol{\alpha}}$ computed in the previous iteration. In analogy to the covariance square root filter, a Householder transformation has also been proposed for performing the time propagation, but it has been shown to be equivalent to, but less efficient than, the Potter-type algorithm given in (7-64) [4].

Thus, in the inverse covariance square root filter, measurement updates are conducted in vector form through a triangularization procedure, and time propagations involve iterative applications of a Potter-type scalar incorporation algorithm. This is in direct opposition to the covariance square root filter. As a result, its time propagations are more efficient than those of the covariance square root filter for the typical case in which the state dimension is much greater than the dynamic noise dimension s. On the other hand, its measurement update is more efficient only when the measurement dimension m is considerably greater than n. Alternative, efficient forms of this filter, also known as square root information filters, have been developed and used extensively for certain applications [5,8]. Although most applications have shown the covariance square root filter to be more efficient computationally, there are circumstances $(m \gg n \gg s)$ under which the inverse covariance square root formulation is preferable. Section 7.8 will compare the various forms explicitly.

7.7 U–D COVARIANCE FACTORIZATION FILTER

Another approach to enhancing the numerical characteristics of the optimal filter algorithm is known as "U–D covariance factorization," developed by Bierman and Thornton [1, 6–8, 14–17, 29, 30]. Rather than decomposing the covariance into its square root factors as in (7-2) and (7-3), this method expresses the covariances before and after measurement incoporation as

$$\mathbf{P}(t_i^-) = \mathbf{U}(t_i^-)\mathbf{D}(t_i^-)\mathbf{U}^T(t_i^-) \tag{7-65}$$

$$\mathbf{P}(t_i^+) = \mathbf{U}(t_i^+)\mathbf{D}(t_i^+)\mathbf{U}^T(t_i^+) \tag{7-66}$$

where the $\mathbf{U}$ matrices are upper triangular and unitary (with ones along the diagonal) and the $\mathbf{D}$ matrices are diagonal. Although covariance square roots are never explicitly evaluated in this method, this filter algorithm is included in this chapter because (1) $\mathbf{U}\mathbf{D}^{1/2}$ corresponds directly to the covariance square root of the Carlson filter in Section 7.5, and the Carlson filter in fact partially motivated this filter development, and (2) the U–D covariance factorization filter shares the advantages of the square root filters discussed previously: guaranteeing nonnegativity of the computed covariance and being numerically accurate and stable. (Merely being a square root filter is not a sufficient condition for numerical accuracy and stability, but the algorithms discussed previously do have these attributes.) Like the Carlson filter, triangular forms are maintained so that this algorithm is considerably more efficient in terms of computations and storage than the Potter filter. Though similar in concept and computation to the Carlson filter, this algorithm does not require any of the $(nm + s)$ computationally expensive scalar square roots as processed in the former.

Before considering the filter algorithm itself, let us demonstrate that, given some $\mathbf{P}$ as an n-by-n symmetric, positive semidefinite matrix, a unit upper triangular factor $\mathbf{U}$ and diagonal factor $\mathbf{D}$ such that $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ can always be generated. Although such $\mathbf{U}$ and $\mathbf{D}$ matrices are not unique, a uniquely defined pair can in fact be generated through an algorithm closely related to the backward running Cholesky decomposition algorithm, (7-7). This will be shown by explicitly displaying the result. First, for the nth column

$$D_{nn} = P_{nn}$$

$$U_{in} = \begin{cases} 1 & i = n \\ P_{in}/D_{nn} & i = n-1, n-2, \dots, 1 \end{cases} \tag{7-67a}$$

Then for the remaining columns, for $j = n-1, n-2, \dots, 1$, compute

$$D_{jj} = P_{jj} - \sum_{k=j+1}^{n} D_{kk}U_{jk}^2$$

$$U_{ij} = \begin{cases} 0 & i > j \\ 1 & i = j \\ [P_{ij} - \sum_{k=j+1}^{n} D_{kk}U_{ik}U_{jk}]/D_{jj} & i = j-1, j-2, \dots, 1 \end{cases} \tag{7-67b}$$

This is useful for defining the required **U–D** factors of $\mathbf{P}_0$ and the $\mathbf{Q}_d$ time history for a given application.

To develop the filter algorithm itself, first consider a scalar measurement update, for which $\mathbf{H}(t_i)$ is 1-by-n. For convenience, we drop the time index and let $\mathbf{P}(t_i^-) = \mathbf{P}^-$, $\mathbf{P}(t_i^+) = \mathbf{P}^+$, and so forth. The Kalman update

$$\mathbf{P}^+ = \mathbf{P}^- - (\mathbf{P}^-\mathbf{H}^\mathrm{T})(1/a)(\mathbf{H}\mathbf{P}^-), \qquad a = \mathbf{H}\mathbf{P}^-\mathbf{H}^\mathrm{T} + R \tag{7-68}$$

can be factored as

$$\mathbf{U}^+\mathbf{D}^+\mathbf{U}^{+\mathrm{T}} = \mathbf{U}^-\mathbf{D}^-\mathbf{U}^{-\mathrm{T}} - (1/a)(\mathbf{U}^-\mathbf{D}^-\mathbf{U}^{-\mathrm{T}}\mathbf{H}^\mathrm{T})\mathbf{H}\mathbf{U}^-\mathbf{D}^-\mathbf{U}^{-\mathrm{T}}$$

$$= \mathbf{U}^-[\mathbf{D}^- - (1/a)(\mathbf{D}^-\mathbf{U}^{-\mathrm{T}}\mathbf{H}^\mathrm{T})(\mathbf{D}^-\mathbf{U}^{-\mathrm{T}}\mathbf{H}^\mathrm{T})^\mathrm{T}]\mathbf{U}^{-\mathrm{T}} \tag{7-69}$$

Note that $\mathbf{U}^{-\mathrm{T}}$ is $(\mathbf{U}^-)^\mathrm{T}$, as distinct from $\mathbf{S}^{-\mathrm{T}} = (\mathbf{S}^{-1})^\mathrm{T}$ in the previous section. Defining the n-vectors $\mathbf{f}$ and $\mathbf{v}$ as

$$\mathbf{f} = \mathbf{U}^{-\mathrm{T}}\mathbf{H}^\mathrm{T} \tag{7-70a}$$

$$\mathbf{v} = \mathbf{D}^-\mathbf{f}; \qquad \text{i.e.,} \quad v_j = D_{jj}^- f_j, \qquad j = 1, 2, \ldots, n \tag{7-70b}$$

and substituting into (7-69) yields

$$\mathbf{U}^+\mathbf{D}^+\mathbf{U}^{+\mathrm{T}} = \mathbf{U}^-[\mathbf{D}^- - (1/a)\mathbf{v}\mathbf{v}^\mathrm{T}]\mathbf{U}^{-\mathrm{T}} \tag{7-71}$$

Now let $\bar{\mathbf{U}}$ and $\bar{\mathbf{D}}$ be the U–D factors of $[\mathbf{D}^- - (1/a)\mathbf{v}\mathbf{v}^\mathrm{T}]$:

$$\bar{\mathbf{U}}\bar{\mathbf{D}}\bar{\mathbf{U}}^\mathrm{T} = [\mathbf{D}^- - (1/a)\mathbf{v}\mathbf{v}^\mathrm{T}] \tag{7-72}$$

so that (7-71) can be written as

$$\mathbf{U}^+\mathbf{D}^+\mathbf{U}^{+\mathrm{T}} = [\mathbf{U}^-\bar{\mathbf{U}}]\bar{\mathbf{D}}[\mathbf{U}^-\bar{\mathbf{U}}]^\mathrm{T} \tag{7-73}$$

Since $\mathbf{U}^-$ and $\bar{\mathbf{U}}$ are unit upper triangular, this then yields

$$\mathbf{U}^+ = \mathbf{U}^-\bar{\mathbf{U}} \tag{7-74a}$$

$$\mathbf{D}^+ = \bar{\mathbf{D}} \tag{7-74b}$$

In this manner, the problem of factoring the Kalman filter measurement update has been reduced to the problem of factoring a symmetric matrix, $[\mathbf{D} - (1/a)\mathbf{v}\mathbf{v}^\mathrm{T}]$ into $\bar{\mathbf{U}}$ and $\bar{\mathbf{D}} = \mathbf{D}^+$. These factors can be generated [6, 14–16] recursively by letting $a_0 = R$ and computing, for $j = 1, 2, \ldots, n$,

$$a_j = \sum_{k=1}^{j} D_{kk} f_k^2 + R$$

$$\bar{D}_{jj} = D_{jj}a_{j-1}/a_j$$

$$\bar{U}_{ij} = \begin{cases} -D_{ii}f_i f_j/a_{j-1} & i = 1, 2, \ldots, j-1 \\ 1 & i = j \\ 0 & i = j+1, j+2, \ldots, n \end{cases} \tag{7-75}$$

Thus, $[\mathbf{D} - (1/a)\mathbf{v}\mathbf{v}^T]$ is scanned and $\bar{\mathbf{U}}$ is generated column by column, as depicted in Fig. 7.3. The validity of the terms generated in (7-75) can be demonstrated by substituting them into (7-67) and showing that the resulting $[P_{ij}]$ matrix is in fact $[\mathbf{D} - (1/a)\mathbf{v}\mathbf{v}^T]$.

The *scalar measurement update for the* **U–D** *covariance factorization filter* can now be specified. At time t_i, $\mathbf{U}(t_i^-)$ and $\mathbf{D}(t_i^-)$ are available from a previous time propagation (to be discussed). Using the measurement value z_i and the known 1-by-n $\mathbf{H}(t_i)$ and scalar $R(t_i)$, one computes

$$\begin{aligned}
\mathbf{f} &= \mathbf{U}^T(t_i^-)\mathbf{H}^T(t_i) \\
v_j &= D_{jj}(t_i^-)f_j \qquad j = 1, 2, \ldots, n \\
a_0 &= R
\end{aligned} \tag{7-76}$$

Then, for $k = 1, 2, \ldots, n$, calculate the results of (7-75), but in a more efficient manner as

$$\left.\begin{aligned}
a_k &= a_{k-1} + f_k v_k \\
D_{kk}(t_i^+) &= D_{kk}(t_i^-)\, a_{k-1}/a_k \\
b_k &\leftarrow v_k \\
p_k &= -f_k/a_{k-1} \\
U_{jk}(t_i^+) &= U_{jk}(t_i^-) + b_j p_k \\
b_j &\leftarrow b_j + U_{jk}(t_i^-)v_k
\end{aligned}\quad\right\} \quad j = 1, 2, \ldots, (k-1) \tag{7-77}$$
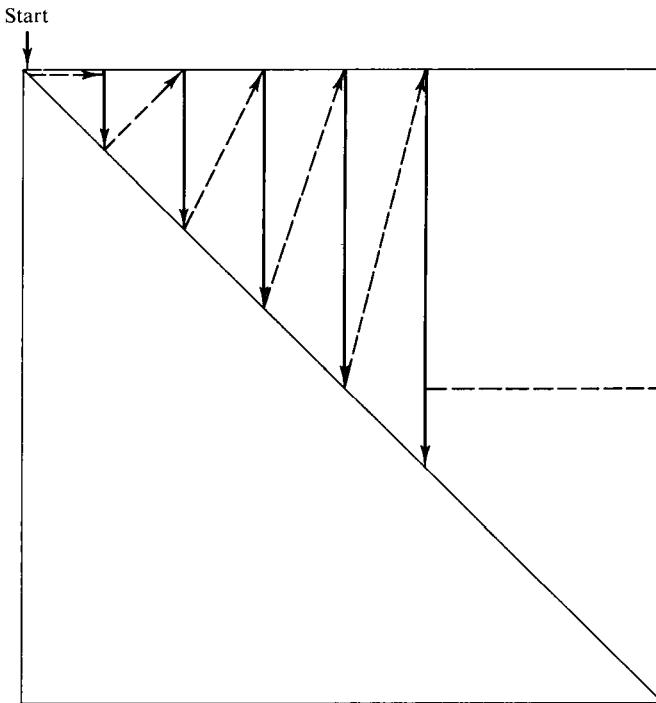
In (7-77), $\leftarrow$ denotes replacement, exploiting the technique of "writing over" old variables for efficiency. For $k = 1$, only the first three equations need be processed. After the n iterations of (7-77), $\mathbf{U}(t_i^+)$ and $\mathbf{D}(t_i^+)$ have been computed, and the filter gain $\mathbf{K}(t_i)$ can be calculated in terms of the n-vector $\mathbf{b}$ made up of components $b_1, b_2, \ldots, b_n$ computed in the last iteration of (7-77), and the state updated as

$$\begin{aligned}
\mathbf{K}(t_i) &= \mathbf{b}/a_n \\
\hat{\mathbf{x}}(t_i^+) &= \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]
\end{aligned} \tag{7-78}$$

Vector measurement updates would be performed component by component, requiring a transformation of variables as in Section 7.4 if $\mathbf{R}(t_i)$ is not originally diagonal.

EXAMPLE 7.8 Consider the same problem treated previously, such that

$$\mathbf{P}(t_i^-) = \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix}, \qquad \mathbf{H}(t_i) = [\tfrac{1}{3} \ \ 1], \qquad R(t_i) = 4$$

Start

$$[\mathbf{D} - (1/a)\mathbf{v}\mathbf{v}^{\mathrm{T}}] \text{ and } \bar{\mathbf{U}}$$

FIG. 7.3   Scanning of $[\mathbf{D} - (1/a)\mathbf{v}\mathbf{v}^{\mathrm{T}}]$ and generation of $\bar{\mathbf{U}}$.

The factors of $\mathbf{P}(t_i^{-})$ are obtained from (7-67) as

$$D_{22}^{-} = P_{22}^{-} = 13$$

$$U_{22}^{-} = 1, \qquad U_{12}^{-} = P_{12}^{-}/D_{22}^{-} = 9/13$$

$$D_{11}^{-} = P_{11}^{-} - D_{22}^{-}U_{12}^{-2} = 9 - 13(9^2/13^2) = 36/13$$

$$U_{21}^{-} = 0, \qquad U_{11}^{-} = 1$$

Thus

$$\mathbf{U}(t_i^{-}) = \begin{bmatrix} 1 & 9/13 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{D}(t_i^{-}) = \begin{bmatrix} 36/13 & 0 \\ 0 & 13 \end{bmatrix}$$

Initialization by (7-76) yields

$$\mathbf{f} = \begin{bmatrix} 1 & 0 \\ 9/13 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 16/13 \end{bmatrix}$$

$$v_1 = [36/13][1/3] = 12/13$$

$$v_2 = [13][16/13] = 16$$

$$a_0 = 4$$

The first iteration of (7-77) produces

$$a_1 = 4 + [1/3][12/13] = 56/13$$
$$D_{11}(t_i^+) = [36/13][4]/[56/13] = 18/7$$
$$b_1 \leftarrow 12/13$$

The second iteration yields

$$a_2 = [56/13] + [16/13][16] = 24$$
$$D_{22}(t_i^+) = [13][56/13]/[24] = 7/3$$
$$b_2 \leftarrow 16$$
$$p_2 = -[16/13]/[56/13] = -2/7$$
$$U_{12}(t_i^+) = [9/13] + [12/13][-2/7] = 3/7$$
$$b_1 \leftarrow [12/13] + [9/13][16] = 12$$

Finally, (7-78) generates

$$\mathbf{K}(t_i) = \begin{bmatrix} 12 \\ 16 \end{bmatrix} \Big/ 24 = \begin{bmatrix} 1/2 \\ 2/3 \end{bmatrix}$$

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)[z_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)]$$

Note that the gain $\mathbf{K}(t_i)$ agrees with previous results and that

$$\mathbf{U}(t_i^+)\mathbf{D}(t_i^+)\mathbf{U}^{\mathrm{T}}(t_i^+) = \begin{bmatrix} 1 & 3/7 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 18/7 & 0 \\ 0 & 7/3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 3/7 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 7/3 \end{bmatrix}$$

which is also consistent with the earlier computations.   ∎

The time propagation of the $\mathbf{U}-\mathbf{D}$ factors employs a generalized Gram–Schmidt orthogonalization to preserve numerical accuracy while attaining computational efficiency [29]. Given the covariance time propagation relation

$$\mathbf{P}(t_{i+1}^-) = \mathbf{\Phi}(t_{i+1}, t_i)\mathbf{P}(t_i^+)\mathbf{\Phi}^{\mathrm{T}}(t_{i+1}, t_i) + \mathbf{G}_{\mathrm{d}}(t_i)\mathbf{Q}_{\mathrm{d}}(t_i)\mathbf{G}_{\mathrm{d}}^{\mathrm{T}}(t_i) \qquad (7\text{-}79)$$

and the $\mathbf{U}-\mathbf{D}$ factors of $\mathbf{P}(t_i^+)$, we desire the factors $\mathbf{U}(t_{i+1}^-)$ and $\mathbf{D}(t_{i+1}^-)$ such that $[\mathbf{U}(t_{i+1}^-)\mathbf{D}(t_{i+1}^-)\mathbf{U}^{\mathrm{T}}(t_{i+1}^-)]$ equals the right hand side of (7-79). Without loss of generality, $\mathbf{Q}_{\mathrm{d}}(t_i)$ is assumed diagonal, since, given the n-by-n matrix $[\mathbf{G}_{\mathrm{d}}(t_i)\mathbf{Q}_{\mathrm{d}}(t_i)\mathbf{G}_{\mathrm{d}}^{\mathrm{T}}(t_i)]$, (7-67) can be used to generate $\mathbf{G}_{\mathrm{d}}(t_i)$ as its $\mathbf{U}$-factor and $\mathbf{Q}_{\mathrm{d}}(t_i)$ as its $\mathbf{D}$-factor.

If an n-by-$(n + s)$ matrix $\mathbf{Y}(t_{i+1}^-)$ and an $(n + s)$-by-$(n + s)$ diagonal matrix $\tilde{\mathbf{D}}(t_{i+1}^-)$ are defined as

$$\mathbf{Y}(t_{i+1}^-) = [\mathbf{\Phi}(t_{i+1}, t_i)\mathbf{U}(t_i^+) \mid \mathbf{G}_{\mathrm{d}}(t_i)] \qquad (7\text{-}80a)$$

$$\tilde{\mathbf{D}}(t_{i+1}^-) = \begin{bmatrix} \mathbf{D}(t_i^+) & \vdots & \mathbf{0} \\ \cdots\cdots & + & \cdots\cdots \\ \mathbf{0} & \vdots & \mathbf{Q}_{\mathrm{d}}(t_i) \end{bmatrix} \qquad (7\text{-}80b)$$

then it can be seen that $[\mathbf{Y}(t_{i+1}^-)\tilde{\mathbf{D}}(t_{i+1}^-)\mathbf{Y}^{\mathrm{T}}(t_{i+1}^-)]$ satisfies (7-79). Similar to the development of (7-29)–(7-55) of Section 7.5, the desired result can be generated through a Gram–Schmidt procedure applied to $\mathbf{Y}(t_{i+1}^-)$. The only significant

modification is that the inner products used in the procedure are weighted inner products: whereas in (7-31) the inner product of $\tilde{s}^j$ [a column of $\tilde{S}^T(t_{i+1}^-)$] and a basis vector b^k was written as $[\tilde{s}^{jT}b^k]$, here the inner product of y^j [a column of $Y^T(t_{i+1}^-)$] and a basis vector b^k would be written as $[y^{jT}\tilde{D}(t_{i+1}^-)b^k]$. When an analogous development is made, $D(t_i^+)$ and $U(t_i^+)$ can be identified as, for $j = 1, 2, \ldots, n$ and $k = j, j+1, \ldots, n$,

$$D_{jj}(t_{i+1}^-) = [b^j]^T\tilde{D}(t_{i+1}^-)b^j \tag{7-81a}$$

$$U_{jk}(t_{i+1}^-) = \frac{1}{D_{kk}(t_{i+1}^-)}\{[y^j]^T\tilde{D}(t_{i+1}^-)b^k\} \tag{7-81b}$$

As in Section 7.5, the actual computational algorithm is the efficient, numerically superior modified weighted Gram–Schmidt (MWGS) method. Thus, the *time propagation relations* are to compute $Y(t_{i+1}^-)$ and $\tilde{D}(t_{i+1}^-)$ as in (7-80), and initialize n vectors, each of dimension $(n + s)$, through

$$[a_1 \quad a_2 \mid \ldots \mid a_n] = Y^T(t_{i+1}^-) \tag{7-82}$$

and then to iterate on the following relations for $k = n, n-1, \ldots, 1$:

$$c_k = \tilde{D}(t_{i+1}^-)a_k \qquad (c_{kj} = \tilde{D}_{jj}(t_{i+1}^-)a_{kj}, \quad j = 1, 2, \ldots, n)$$
$$D_{kk}(t_{i+1}^-) = a_k^T c_k$$
$$d_k = c_k/D_{kk}(t_{i+1}^-) \tag{7-83}$$
$$\left.\begin{array}{l} U_{jk}(t_{i+1}^-) = a_j^T d_k \\ a_j \leftarrow a_j - U_{jk}(t_{i+1}^-)a_k \end{array}\right\} \quad j = 1, 2, \ldots, k-1$$

As before, $\leftarrow$ denotes replacement, or "writing over" old variables to reduce storage requirements. On the last iteration, for $k = 1$, only the first two relations need be computed. The state estimate is given by

$$\hat{x}(t_{i+1}^-) = \Phi(t_{i+1}, t_i)\hat{x}(t_i^+) \tag{7-84}$$

EXAMPLE 7.9   Consider the same time propagation as in Examples 7.3, 7.4, and 7.5; let

$$[\Phi P(t_i^+)\Phi^T] = [\Phi S(t_i^+)][S^T(t_i^+)\Phi^T] = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 8 & 8 \\ 8 & 10 \end{bmatrix}$$

$$G_d Q_d G_d^T = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

For the sake of this example, let

$$\Phi = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{bmatrix}, \qquad P(t_i^+) = \begin{bmatrix} 8 & 4 \\ 4 & 4 \end{bmatrix}$$

so that $[\Phi P(t_i^+)\Phi^T]$ is as given above. The U–D factors of $P(t_i^+)$ would be given by a previous measurement update; for this problem, they can be computed from (7-67) as

$$U(t_i^+) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad D(t_i^+) = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

Finally, since $\mathbf{Q}_d$ is assumed to be diagonal, $[\mathbf{G}_d\mathbf{Q}_d\mathbf{G}_d^T]$ can be factored by (7-67) into

$$\mathbf{G}_d = \begin{bmatrix} 1 & \frac{1}{3} \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{Q}_d = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & 3 \end{bmatrix}$$

The time propagation computations are initialized by (7-80) as

$$\mathbf{Y}(t_{i+1}^-) = \left[\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \,\middle|\, \begin{bmatrix} 1 & \frac{1}{3} \\ 0 & 1 \end{bmatrix}\right] = \begin{bmatrix} 1 & 1 & 1 & \frac{1}{3} \\ \frac{1}{2} & \frac{3}{2} & 0 & 1 \end{bmatrix}$$

$$\tilde{\mathbf{D}}(t_{i+1}^-) = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

so that (7-82) yields

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \frac{1}{3} \end{bmatrix}, \qquad \mathbf{a}_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{3}{2} \\ 0 \\ 1 \end{bmatrix}$$

The first iteration of (7-83), for $k = n = 2$, produces

$$\mathbf{c}_2 = \begin{bmatrix} 4 \cdot \frac{1}{2} \\ 4 \cdot \frac{3}{2} \\ \frac{2}{3} \cdot 0 \\ 3 \cdot 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 0 \\ 3 \end{bmatrix}$$

$$D_{22}(t_{i+1}^-) = \begin{bmatrix} \frac{1}{2} & \frac{3}{2} & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 6 \\ 0 \\ 3 \end{bmatrix} = 13$$

$$\mathbf{d}_2 = \begin{bmatrix} 2 \\ 6 \\ 0 \\ 3 \end{bmatrix} \cdot \frac{1}{13} = \begin{bmatrix} 2/13 \\ 6/13 \\ 0 \\ 3/13 \end{bmatrix}$$

$$U_{12}(t_{i+1}^-) = \begin{bmatrix} 1 & 1 & 1 & 1/3 \end{bmatrix}\begin{bmatrix} 2/13 \\ 6/13 \\ 0 \\ 3/13 \end{bmatrix} = 9/13$$

$$\mathbf{a}_1 \leftarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1/3 \end{bmatrix} - \frac{9}{13}\begin{bmatrix} 1/2 \\ 3/2 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 17/26 \\ -1/26 \\ 1 \\ -14/39 \end{bmatrix}$$

The second iteration, for $k = 1$, generates

$$\mathbf{c}_1 = \begin{bmatrix} 4 & \cdot & 17/26 \\ 4 & \cdot & -1/26 \\ 2/3 & \cdot & 1 \\ 3 & \cdot & -14/39 \end{bmatrix} = \begin{bmatrix} 34/13 \\ -2/13 \\ 2/3 \\ -14/13 \end{bmatrix}$$

RADCLIFFF

$$D_{11}(t_{i+1}^-) = \begin{bmatrix} 17/26 & -1/26 & 1 & -14/39 \end{bmatrix} \begin{bmatrix} 34/13 \\ -2/13 \\ 2/3 \\ -14/13 \end{bmatrix} = 36/13$$

Thus, $\mathbf{U}(t_{i+1}^-)$ and $\mathbf{D}(t_{i+1}^-)$ have been generated as

$$\mathbf{U}(t_{i+1}^-) = \begin{bmatrix} 1 & 9/13 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{D}(t_{i+1}^-) = \begin{bmatrix} 36/13 & 0 \\ 0 & 13 \end{bmatrix}$$

Note that

$$\mathbf{U}(t_{i+1}^-)\mathbf{D}(t_{i+1}^-)\mathbf{U}^T(t_{i+1}^-) = \begin{bmatrix} 9 & 9 \\ 9 & 13 \end{bmatrix}$$

as found in the earlier examples or by adding $[\mathbf{\Phi P}(t_i^+)\mathbf{\Phi}^T]$ and $[\mathbf{G_d Q_d G_d}^T]$ directly. ∎

7.8   FILTER PERFORMANCE AND REQUIREMENTS

The algorithms of this chapter have been investigated in order to implement the optimal filtering solution given by the Kalman filter, but without the numerical instability and inaccuracy of that algorithm when processed with finite wordlength. In this section, both the numerical advantages and the increased computational burden of these filters will be delineated.

An algorithm can be said to be numerically stable if the computed result of the algorithm corresponds to an exactly computed solution to a problem that is only slightly perturbed from the original one [31]. By this criterion, the *Kalman filter is numerically unstable* [6], in both the conventional and Joseph formulations. In contrast, *all of the filters described in this chapter can be shown to be numerically stable*.

The numerical conditioning of a set of computations can be described in part by what is called a "condition number," a concept which is often used to analyze the effects of perturbations in linear equations. If $\mathbf{A}$ is a matrix, not necessarily square, then the condition number $k(\mathbf{A})$ associated with $\mathbf{A}$ is defined by [22]:

$$k(\mathbf{A}) = \sigma_{max}/\sigma_{min} \tag{7-85}$$

where σ_{max}^2 and σ_{min}^2 are the maximum and minimum eigenvalues of $\mathbf{A}^T\mathbf{A}$, respectively. When computing in base 10 (or base 2) arithmetic with N significant digits (or bits), numerical difficulties may be expected as $k(\mathbf{A})$ approaches 10^N

(or 2^N). For instance, if the maximum and minimum numbers of interest, σ_{max} and σ_{min}, were 100000 and 000001 (in base 10 or 2), then to add these values together and obtain 100001 without numerical difficulties would require at least six significant figures (digits or bits). But,

$$k(\mathbf{P}) = k(\mathbf{SS}^T) = [k(\mathbf{S})]^2 \tag{7-86}$$

Therefore, while numerical operations on the covariance $\mathbf{P}$ may encounter difficulties when $k(\mathbf{P}) = 10^N$ (or 2^N), those same numerical problems would arise when $k(\mathbf{S}) = 10^{N/2}$ (or $2^{N/2}$) according to (7-86): *the same numerical precision is achieved with half the wordlength.*

EXAMPLE 7.10   This example and the next illustrate the improved numerical characteristics of the square root filters. To simulate roundoff, let $e \ll 1$ be such that

$$1 + e \overset{r}{\neq} 1$$
$$1 + e^2 \overset{r}{=} 1$$

where $\overset{r}{=}$ means equal due to rounding. Consider a scalar measurement update of a two-state problem, with

$$\mathbf{P}(t_i^-) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{H}(t_i) = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad R(t_i) = e^2$$

and compare the computed results of the filters of Chapter 5 and of this chapter. Note that

$$\mathbf{P}(t_i^-) = \mathbf{P}^{-1}(t_i^-) = \mathbf{S}(t_i^-) = \mathbf{S}^{-1}(t_i^-) = \mathbf{U}(t_i^-) = \mathbf{D}(t_i^-) = \mathbf{I}$$

and that the exact covariance $\mathbf{P}(t_i^+)$ for this example is:

$$\mathbf{P}(t_i^+) = \begin{bmatrix} e^2/(1 + e^2) & 0 \\ 0 & 1 \end{bmatrix}$$

The computed results are
  (a)   conventional Kalman

$$\mathbf{P}(t_i^+) \overset{r}{=} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

  (b)   Joseph form, Kalman

$$\mathbf{P}(t_i^+) \overset{r}{=} \begin{bmatrix} e^2 & 0 \\ 0 & 1 \end{bmatrix}$$

  (c)   Potter covariance square root

$$\mathbf{S}(t_i^+) \overset{r}{=} \begin{bmatrix} e & 0 \\ 0 & 1 \end{bmatrix}$$

  (d)   Carlson covariance square root

$$\mathbf{S}(t_i^+) \overset{r}{=} \begin{bmatrix} e & 0 \\ 0 & 1 \end{bmatrix}$$

(e)   inverse covariance

$$\mathbf{P}^{-1}(t_i{}^+) \overset{r}{=} \begin{bmatrix} 1/e^2 & 0 \\ 0 & 1 \end{bmatrix}$$

(f)   inverse covariance square root

$$\mathbf{S}^{-1}(t_i{}^+) \overset{r}{=} \begin{bmatrix} 1/e & 0 \\ 0 & 1 \end{bmatrix}$$

(g)   U–D factor

$$\mathbf{U}(t_i{}^+) \overset{r}{=} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{D}(t_i{}^+) \overset{r}{=} \begin{bmatrix} e^2 & 0 \\ 0 & 1 \end{bmatrix}$$

For this example, all but the conventional Kalman filter yield nonsingular and nearly exact answers. Although the difference between 0 and e^2 in the upper left element of $\mathbf{P}(t_i{}^+)$ may seem insignificant, it can have grave consequences. For instance, assume no dynamics and let a second measurement of the same form be taken. The gain $\mathbf{K}$ computed by the conventional Kalman filter would be

$$\mathbf{K}(t_i{}^{++}) = \mathbf{P}(t_i{}^+)\mathbf{H}(t_i)/[\mathbf{H}(t_i)\mathbf{P}(t_i{}^+)\mathbf{H}^{\mathrm{T}}(t_i) + R(t_i)]$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix}\bigg/ [0 + 1] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

whereas the correct value is

$$\mathbf{K}(t_i{}^{++}) = \frac{e^2}{1 + e^2}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\bigg/ \left\{\frac{e^2}{1 + e^2} + e^2\right\} \cong \frac{1}{2}\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

as would be calculated correctly by ·the Joseph form in this case.   ∎

EXAMPLE 7.11   Consider the same problem as in Example 7.10, but let $\mathbf{H}(t_i)$ now be $[1 \; 1]$ instead of $[1 \; 0]$. In this case, the exact answer is

$$\mathbf{P}(t_i{}^+) = \frac{1}{2 + e^2}\begin{bmatrix} 1 + e^2 & -1 \\ -1 & 1 + e^2 \end{bmatrix}$$

The computed results are
  (a)   conventional Kalman

$$\mathbf{P}(t_i{}^+) \overset{r}{=} \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

(b)   Joseph form Kalman

$$\mathbf{P}(t_i{}^+) \overset{r}{=} \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

(c)   Potter covariance square root

$$\mathbf{S}(t_i{}^+) \overset{r}{=} \begin{bmatrix} 1 + e/\sqrt{2} & 0 \\ -1 + e/\sqrt{2} & 1 + e/\sqrt{2} \end{bmatrix}$$

(d)   Carlson covariance square root

$$\mathbf{S}(t_i{}^+) \stackrel{\mathrm{r}}{=} \begin{bmatrix} e & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{bmatrix}$$

(e)   inverse covariance

$$\mathbf{P}^{-1}(t_i{}^+) \stackrel{\mathrm{r}}{=} \frac{1}{e^2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

(f)   inverse covariance square root

$$\mathbf{S}^{-1}(t_i{}^+) \stackrel{\mathrm{r}}{=} \frac{1}{e} \begin{bmatrix} -1 & -1 \\ 0 & e\sqrt{2} \end{bmatrix}$$

(g)   **U–D** factor

$$\mathbf{U}(t_i{}^+) \stackrel{\mathrm{r}}{=} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{D}(t_i{}^+) \stackrel{\mathrm{r}}{=} \begin{bmatrix} e^2 & 0 \\ 0 & \tfrac{1}{2} \end{bmatrix}$$

In this case, only the square root and **U–D** implementations yield nonsingular results. Such singular $\mathbf{P}(t_i{}^+)$ or $\mathbf{P}^{-1}(t_i{}^+)$ matrices would again yield a zero gain $\mathbf{K}$ if a second measurement of the same form were processed, while the square root and **U–D** factor filters compute a gain which is nearly exact. Moreover, even though $\mathbf{S}$, $\mathbf{S}^{-1}$, $\mathbf{U}$, and $\mathbf{D}$ are nonsingular, the associated value of $\mathbf{P}(t_i{}^+)$ or $\mathbf{P}^{-1}(t_i{}^+)$ found by multiplication would be rounded to a singular matrix; thus it is better not to perform such computations explicitly, and the time propagations based on triangularization are to be preferred over the RSS method which performs such multiplication.   ∎

The improved numerical characteristics of the square root and U–D factorization filters are achieved at the expense of increased computational burden. Letting n be the dimension of the state vector **x**, s be the dimension of the dynamic driving noise **w**, and m be the dimension of the measurement **z** and its corruptive noise **v**, we now determine the number of mathematical operations required by the various filters, assuming that

(1)   all implementations take advantage of symmetry and zeros as they appear in general forms,
(2) $\mathbf{R}(t_i)$ and $\mathbf{Q}_\mathrm{d}(t_i)$ are diagonal, and
(3)   the inverse covariance and inverse covariance square root filters generate explicit state estimates, $\hat{\mathbf{x}}(t_i{}^-)$ and $\hat{\mathbf{x}}(t_i{}^+)$.

Table 7.1 presents the number of operations for one time propagation and one measurement update required by

(1)   Kalman filter—with conventional and Joseph form measurement update,
(2)   Potter covariance square root filter—with MGS and Householder time propagations,
(3)   Carlson covariance square root filter—with matrix RSS and MGS time propagations,
(4)   inverse covariance filter [using (5-91) for time propagation],

TABLE 7.1

Operations Required for One Time Propagation and One Measurement Update

Filter	Adds (all times $\frac{1}{6}$)	Multiplies (all times $\frac{1}{6}$)	Divides	Square roots
Conventional Kalman	$9n^3 + 3n^2(3m + s - 1)$ $+ n(15m + 3s - 6)$	$9n^3 + 3n^2(3m + s + 3)$ $+ n(27m + 9s)$	m	0
Joseph form Kalman	$18n^3 + 3n^2(5m + s - 10)$ $+ n(9m^2 + 6m + 3s)$ $+ 3m^3 - 6m^2 + 3m$	$18n^3 + 3n^2(5m + s + 4)$ $+ n(9m^2 + 24m + 9s)$ $+ 3m^3 + 9m^2 - 6m$	$2m - 1$	0
Potter covariance square root (MGS)	$12n^3 + 3n^2(6m + 2s)$ $+ n(6m - 6) + 6m$	$12n^3 + 3n^2(6m + 2s + 2)$ $+ n(24m + 6s) + 12m$	$n + 2m$	$n + m$
Potter covariance square root (Householder)	$10n^3 + 3n^2(6m + 2s - 1)$ $+ n(6m + 5) + 6m$	$10n^3 + 3n^2(6m + 2s + 2)$ $+ n(24m + 6s + 8) + 12m$	$n + 2m$	$n + m$
Carlson covariance square root (RSS)	$5n^3 + 3n^2(3m + s + 1)$ $+ n(9m + 3s - 14)$ $+ 2s^3 + 4s$	$5n^3 + 3n^2(4m + s + 3)$ $+ n(30m + 9s - 2)$ $+ 2s^3 + 6s^2 - 2s$	$2mn + s$	$mn + s$
Carlson covariance square root (MGS)	$9n^3 + 3n^2(3m + s - 1)$ $+ 3n(3m + 3s - 8)$ $+ 2s^3 + 6s^2 + 4s$	$9n^3 + 3n^2(4m + s + 2)$ $+ 3n(10m + 5s - 7)$ $+ 2s^3 + 12s^2 + 4s$	$2mn + s$	$mn + s$
Inverse covariance	$10n^3 + 3n^2(m + 3s + 2)$ $+ n(9m + 9s - 16)$	$10n^3 + 3n^2(m + 3s + 6)$ $+ n(15m + 21s - 10)$	$2s - 1$	0
Inverse covariance square root	$9n^3 + 3n^2(2m + 6s + 5)$ $+ n(12m + 6s - 6)$	$9n^3 + 3n^2(2m + 6s + 6)$ $+ n(12m + 24s + 3) + 6s$	$2n + 2s$	$n + s$
U–D factor	$9n^3 + 3n^2(3m + 2s + 2)$ $+ 3n(3m + 1)$	$9n^3 + 3n^2(3m - 2s + 7)$ $+ 3n(m + 4s - 4) - 6s$	$n(m + 1) - 1$	0

(5)   inverse covariance square root filter (MGS update), and
(6)   **U–D** covariance factorization filter.

This table can be used to project the computation time required by each filter formulation for a given application. Note that if, instead of assuming $\mathbf{Q}_d(t_i)$ to be diagonal, we were to assume that the *n*-by-n $[\mathbf{G}_d(t_i)\mathbf{Q}_d(t_i)\mathbf{G}_d^T(t_i)]$ or the *n*-by-s $[\mathbf{G}_d(t_i)\mathbf{W}_d(t_i)]$ were known, there would be $\frac{1}{2}ns(n + 3)$ fewer multiplies and $\frac{1}{2}n(n + 1)(s - 1)$ fewer adds in filter forms 1, 3 with RSS time propagation, and 4, or *ns* fewer multiplies in forms 2 and 3 with MGS time propagation. Problem 7.13 extends this table to account for taking advantage of matrix sparsity and structure in typical estimation problems.

EXAMPLE 7.12   To put the algebraic expressions of Table 7.1 into perspective, Table 7.2 presents the number of operations required for one time propagation and one measurement update for the case of $n = 10$, $s = 10$, and $m = 2$. The noise dimension s was intentionally set equal

TABLE 7.2

Operations for One Total Filter Recursion[a]

Filter	Adds	Multiplies	Divides	Square roots	Time (msec)
Conventional Kalman	2340	2690	2	0	17.36
Joseph form Kalman	3631	4498	3	0	28.27
Potter covariance square root (MGS)	3612	3884	14	12	26.49
Potter covariance square root (Householder)	3247	3564	14	12	24.19
Carlson covariance square root (RSS)	2080	2560	50	30	18.24
Carlson covariance square root (MGS)	2830	3355	50	30	23.53
Inverse covariance	3520	3950	19	0	25.82
Inverse covariance square root	5080	5455	40	20	37.55
U–D factor	2935	3330	29	0	21.77

[a] $n = s = 10$ and $m = 2$.

to n to correspond to the n-by-n $[G_d(t_i) Q_d(t_i) G_d^T(t_i)]$ being of full rank, typical of an equivalent discrete-time model. The last column in Table 7.2 portrays computer time required for one total filter recursion, neglecting the computations associated with the various subscripting and storage operations for each filter (roughly the same for each), and using single precision instruction times typical of the IBM 360 and some smaller state-of-the-art computers:

$$\text{time for addition} = 2.7 \ \mu\text{sec}$$
$$\text{time for multiplication} = 4.1 \ \mu\text{sec}$$
$$\text{time for division} = 6.6 \ \mu\text{sec}$$
$$\text{time for square root} = 60.0 \ \mu\text{sec}$$

As can be seen from Table 7.2, the covariance square root filters and the U–D covariance factorization filter involve a computational load greater than the conventional Kalman filter, but not so great as to be prohibitive. In fact, the increase is less than that caused by employing the Joseph form of the update equation, which is inferior to these filters in performance. Moreover, since the Kalman filter would probably require double precision operations instead of the single precision assumed to establish Table 7.2, these filters are even more competitive with the Kalman filter than indicated in the table.

Of the square root type filters, the Carlson covariance square root and the U–D covariance factorization filters are the most efficient computationally. The Carlson filter with matrix RSS time propagations requires the least computer time, but this is offset by the degraded numerical accuracy of the matrix RSS method. Thus, the U–D covariance factorization filter would appear to be an exceptionally efficient and numerically advantageous alternative to the conventional Kalman filter for this particular application. ■

7.9   SUMMARY

This chapter presented the concept of square root filters and the closely related U–D covariance factorization filter as viable alternatives to conventional Kalman filters. For a modest increase in computational loading, one obtains optimal filter algorithms equivalent to the Kalman filter if infinite wordlength is assumed, but with vastly superior numerical characteristics with finite word-length. From a numerical analysis standpoint, this is at least as good a solution to troublesome measurement update computations as implementing a Kalman filter in double precision, since the Kalman filter inherently involves unstable numerics.

Of the covariance square root forms, the Carlson filter is more efficient than the Potter form computationally, and it also maintains triangularity of the square root matrices. The U–D covariance factorization filter is comparable to the Carlson filter and does not require square root computations. In comparison, the inverse covariance square root filter is often considerably more burdensome computationally, although it too becomes competitive if the measurement dimension m is very large.

Chandrasekhar-type square root algorithms have also been reported in the literature [25]. However, these have been omitted because they do not appear to be computationally competitive with algorithms presented herein for the nonstationary linear discrete-time estimation problem.

REFERENCES

1. Agee, W. S., and Turner, R. H., "Triangular Decomposition of a Positive Definite Matrix Plus a Symmetric Dyad with Applications to Kalman Filtering," Mathematical Services Branch, Analysis and Computation Division, Tech. Rep. 38, White Sands Missile Range, 1972.
2. Andrews, A., "A Square Root Formulation of the Kalman Covariance Equations," *AIAA J.* **6**(6), 1165–1166 (1968).
3. Bellantoni, J. F., and Dodge, K. W., "A Square Root Formulation of the Kalman–Schmidt Filter," *AIAA J.* **5** (7), 1309–1314 (1967).
4. Bierman, G. J., "A Comparison of Discrete Linear Filtering Algorithms," *IEEE Trans. Aerospace and Electron. Systems* **AES-9** (1) 28–37 (1973).
5. Bierman, G. J., "Sequential Square Root Filtering and Smoothing of Discrete Linear Systems," *Automatica* **10**, 147–158 (1974).
6. Bierman, G. J., "Measurement Updating Using the U–D Factorization," *Proc. IEEE Control and Decision Conf., Houston, Texas* 337–346 (1975).
7. Bierman, G. J., and Thornton, C. L., "Numerical Comparison of Kalman Filter Algorithms: Orbit Determination Case Study," *Automatica* **13**, 23–35 (1977).
8. Bierman, G. J., *Factorization Methods for Discrete Sequential Estimation.* Academic Press, New York, 1977.
9. Björck, A., "Solving Linear Least Squares Problems by Gram–Schmidt Orthogonalization," *BIT* **7**, 1–21 (1967).
10. Businger, P., and Golub, G. H., "Linear Least Squares Solution by Householder Transformations," *Numer. Math.* **7**, 269–276 (1965).

11. Carlson, N. A., "Fast Triangular Formulation of the Square Root Filter," *AIAA J.* **11** (9), 1259–1265 (1973).
12. Dyer, P., and McReynolds, S., "Extension of Square-Root Filtering to Include Process Noise," *J. Optimization Theory Appl.* **3** (6), 444–459 (1969).
13. Fadeeva, V. N., *Computational Methods of Linear Algebra.* Dover, New York, 1959.
14. Fletcher, R., and Powell, M. J. D., "On the Modification of LDL^T Factorizations," *Math. Comp.* **28** (128), 1067–1087 (1974).
15. Gentleman, W. M., "Least Squares Computations by Givens Transformations without Square Roots," *J. Inst. Math. Appl.* **12**, 329–336 (1973).
16. Gill, P. E., Golub, G. H., Murray, W., and Saunders, M. A., "Methods for Modifying Matrix Factorizations," *Math. Comp.* **28** (126), 505–535 (1974).
17. Gill, P. E., Murray, W., and Saunders, M. A., "Methods for Computing and Modifying the LDV Factors of a Matrix," *Math. Comp.* **29** (132), 1051–1077 (1975).
18. Golub, G. H., "Numerical Methods for Solving Linear Least Squares Problems," *Numer. Math.* **7**, 206–216 (1965).
19. Hanson, R. J., and Lawson, C. L., "Extensions and Applications of the Householder Algorithm for Solving Linear Least Squares Problems," *Math. Comp.* **23** (108), 787–812 (1969).
20. Householder, A. S., *The Theory of Matrices in Numerical Analysis.* Blaisdell, Waltham, Massachusetts, 1964.
21. Jordan, T., "Experiments on Error Growth Associated with Some Linear Least Squares Procedures," *Math. Comp.* **22**, 579–588 (1968).
22. Kaminski, P. G., Bryson, A. E. Jr., and Schmidt, S. F., "Discrete Square Root Filtering: A Survey of Current Techniques," *IEEE Trans. Automatic Control* **AC-16** (6), 727–735 (1971).
23. Lawson, C. L., and Hanson, R. J., *Solving Linear Least Squares Problems.* Prentice-Hall, Englewood Cliffs, New Jersey, 1974.
24. Maybeck, P. S., "Solutions to the Kalman Filter Wordlength Problem: Square Root and U–D Covariance Factorizations," Tech. Rep. AFIT-TR-77-6, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, September 1977.
25. Morf, M., and Kailath, T., "Square-Root Algorithms for Least-Squares Estimation," *IEEE Trans. Automatic Control* **AC-20** (4), 487–497 (1975).
26. Potter, J. E., "W Matrix Augmentation," M.I.T. Instrumentation Laboratory Memo SGA 5-64, Cambridge, Massachusetts, January 1964.
27. Rice, J. R., "Experiments on Gram–Schmidt Orthogonalization," *Math. Comp.* **20**, 325–328 (1966).
28. Schmidt, S. F., "Computational Techniques in Kalman Filtering," *in Theory and Applications of Kalman Filtering*, AGARDograph 139, Chapter 3. London, 1970.
29. Thornton, C. L., and Bierman, G. J., "Gram–Schmidt Algorithms for Covariance Propagation," *Proc. IEEE Control and Decision Conf., Houston, Texas* 489–498 (1975).
30. Wampler, R. H., "A Report on the Accuracy of Some Widely Used Least Squares Computer Programs," *J. Amer. Statist. Assoc.* **65** (330), 549–565 (1970).
31. Wilkinson, J. H., *Rounding Errors in Algebraic Processes.* Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

PROBLEMS

7.1 Generate both the lower and upper Cholesky square root matrices of

$$(a) \quad \mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 0.4 \end{bmatrix} \qquad (b) \quad \mathbf{A} = \begin{bmatrix} 3 & 1 & 2 & 1 \\ 1 & 2 & 2 & 1 \\ 2 & 2 & 3 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}$$

7.2 Show the equivalence of Eqs. (7-16) and (7-17) for the Potter measurement update.

7.3 Let $S(t_i^-)$ be given as in Example 7.6, but let the measurement at time t_i be described as

$$\begin{bmatrix} z_1(t_i) \\ z_2(t_i) \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1(t_i) \\ x_2(t_i) \end{bmatrix} + \begin{bmatrix} v_1(t_i) \\ v_2(t_i) \end{bmatrix}$$

with

$$\mathbf{R}(t_i) = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

Show that incorporating $z_2(t_i, \omega_j)$ by a second iteration of the Potter algorithm upon the result of Example 7.6 yields a solution equivalent to the Andrews vector update given by (7-18).

7.4 Consider an application in which a three-state filter is to be updated with two measurements each sample time. Let

$$\mathbf{H} = \begin{bmatrix} 1 & -2 & -1 \\ -1 & -1 & 1 \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}, \qquad \mathbf{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}$$

Explicitly convert this into a form compatible with iterative scalar measurement updating in a Potter covariance square root filter.

7.5 Let a system of interest be described by

$$\begin{bmatrix} x_1(t_{i+1}) \\ x_2(t_{i+1}) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t_i) \\ x_2(t_i) \end{bmatrix} + \begin{bmatrix} w_{d1}(t_i) \\ w_{d2}(t_i) \end{bmatrix}$$

$$\begin{bmatrix} z_1(t_i) \\ z_2(t_i) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1(t_i) \\ x_2(t_i) \end{bmatrix} + \begin{bmatrix} v_1(t_i) \\ v_2(t_i) \end{bmatrix}$$

where the a priori knowledge of $\mathbf{x}(t_0)$ is that it can be modeled as a Gaussian random vector with mean $\hat{\mathbf{x}}_0$ and covariance $\mathbf{P}_0$ given by

$$\hat{\mathbf{x}}_0 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \qquad \mathbf{P}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Let $\mathbf{w}_d(\cdot, \cdot)$ and $\mathbf{v}(\cdot, \cdot)$ be independent white Gaussian noises, each independent of $\mathbf{x}(t_0)$ and of mean zero, and having covariances

$$\mathbf{Q}_d = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

Let $z_1(t_i) = 6$ and $z_2(t_i) = 4$.

Perform the first time propagation from t_0 to t_1, and the update at time t_1, for the Potter covariance square root filter. For the time propagation, use

(a) the matrix root sum squared (RSS) method,
(b) the modified Gram–Schmidt (MGS) technique,
(c) the Householder transformation algorithm.

7.6 Repeat the previous problem, but generate the Carlson filter (using both matrix RSS and MGS time propagations).

7.7 Generate the inverse covariance square root filter for the application given in Problem 7.5. Use both the MGS and Householder measurement updates.

7.8 Generate the U–D covariance factorization filter for the application given in Problem 7.5.

7.9 Generate the U–D factors of the matrices in Problem 7.1.

7.10 You are confronted with an engineer who tells you that he has been investigating square root filters as an alternative to conventional filters. The general problem requires a Gram–Schmidt orthogonalization or a Householder transformation, but the square root covariance filter does not

require such computation for the case of $\mathbf{Q} = \mathbf{0}$, or no dynamic noise. He suggests using very accurate measurements, modeled as essentially perfect, and then using a square root inverse covariance filter. This, he claims, will similarly not require a Gram–Schmidt or Householder algorithm in the filter computations. What is your response to him?

7.11 Explicitly derive the numerical results depicted in Examples 7.10 and 7.11.

7.12 Repeat the calculations of Example 7.12 for

(a) $n = 10, s = 10, m = 5$;

(b) $n = 15, s = 15, m = 2$;

(c) $n = 15, s = 10, m = 2$;

(d) $n = 20, s = 20, m = 2$;

(e) $n = 20, s = 20, m = 5$.

7.13 Often the state variables estimated by a filter can be classified in two general categories: primary system states (as position, velocity, and misalignment error states in navigation filters) and secondary states—for shaping filters to generate outputs affecting either state dynamics or measured outputs. Thus, the state can be partitioned as ($n_1 + n_2 = n$):

$$\mathbf{x}(\cdot,\cdot) = \left[\begin{array}{c} \mathbf{x}_1(\cdot,\cdot) \\ \hline \mathbf{x}_2(\cdot,\cdot) \end{array}\right] \begin{array}{l} \} n_1 \text{ primary states} \\ \} n_2 \text{ secondary states} \end{array}$$

Usually, the propagation of $\mathbf{x}_2(\cdot,\cdot)$ is independent of $\mathbf{x}_1$ so that the state transition matrix $\boldsymbol{\Phi}$ can be partitioned as

$$\boldsymbol{\Phi} = \left[\begin{array}{c|c} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \hline \mathbf{0} & \boldsymbol{\Phi}_{22} \end{array}\right] \begin{array}{l} \} n_1 \\ \} n_2 \end{array}$$

where $\boldsymbol{\Phi}_{11}$ is typically dense, $\boldsymbol{\Phi}_{22}$ is often diagonal, and $\boldsymbol{\Phi}_{12}$ typically contains one or two nonzero elements per column.

(a) Show that if $\mathbf{S}(t_i^+)$ is upper triangular, only the upper left n_1-by-n_1 partition of $[\boldsymbol{\Phi}\mathbf{S}(t_i^+)]$ is nontriangular and requires retriangularization in a Carlson-type filter. (This in fact motivated the choice of *upper* triangular forms for this filter.)

(b) Recalculate the entries of Table 7.1 as a function of n_1 and n_2 instead of n, assuming $\boldsymbol{\Phi}$ to be the form just described.

(c) Repeat the calculations of Example 7.12 for $n_1 = n_2 = 5$, and for the case of $n_1 = 3, n_2 = 7$.

7.14 In Monte Carlo analyses and other types of system simulations, it is often desired to generate samples of a discrete-time white Gaussian noise vector process, described by mean of zero and covariance

$$E\{\mathbf{w}_d(t_i)\mathbf{w}_d^{\mathrm{T}}(t_i)\} = \mathbf{Q}_d(t_i)$$

with $\mathbf{Q}_d(t_i)$ nondiagonal. Independent scalar white Gaussian noises can be simulated readily through use of pseudorandom codes, but the question remains, how does one properly provide for cross-correlations of the scalar noises?

(a) Let $\mathbf{w}_1(\cdot,\cdot)$ be a vector process composed of independent scalar white Gaussian noises of zero mean and unit variance:

$$E\{\mathbf{w}_{1k}(t_i)\} = 0, \qquad E\{\mathbf{w}_{1k}^2(t_i)\} = 1, \qquad k = 1, 2, \ldots, s$$

Show that

$$\mathbf{w}_d(t_i,\cdot) = \begin{cases} \sqrt[r]{\mathbf{Q}_d(t_i)}\,\mathbf{w}_1(t_i,\cdot) \\ \qquad\quad \text{or} \\ \sqrt[c]{\mathbf{Q}_d(t_i)}\,\mathbf{w}_1(t_i,\cdot) \end{cases} \qquad \text{for all } i$$

properly models the desired characteristics.

(b) If $\mathbf{U}_Q(t_i)$ and $\mathbf{D}_Q(t_i)$ are the **U**–**D** factors of $\mathbf{Q}_d(t_i)$, show that

$$\mathbf{w}_d(t_i, \cdot) = \mathbf{U}_Q(t_i)\mathbf{w}_2(t_i, \cdot) \qquad \text{for all } i$$

also provides the desired model if $\mathbf{w}_2(\cdot, \cdot)$ is a vector process composed of independent scalar white Gaussian noises of mean zero and variance

$$E\{\mathbf{w}_{2k}^2(t_i)\} = D_{Qkk}$$

(c) Show the means of simulating $\mathbf{w}_d(\cdot, \cdot)$ if $\mathbf{Q}_d$ is given by

$$\mathbf{Q}_d = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

This page intentionally left blank

Index

411

* Numbers in parentheses are equation numbers.

This page intentionally left blank