

The strip-chart recording shows the phase variation of a very low frequency navigation signal as received at Ames, Iowa, after traversing a 4000-mile path from Hawaii. The originating signal is highly phase stable. However, the received signal exhibits random variations in phase primarily because of variations in the ionosphere and atmospheric noise due to lightning. The physical mechanisms in play in this situation are much too complicated to be described in terms of a deterministic model. Thus, we usually must be content to describe such noiselike waveforms in probabilistic terms. The theoretical description and filtering of noiselike signals is the main theme of this book.

INTRODUCTION TO RANDOM SIGNAL ANALYSIS AND KALMAN FILTERING

ROBERT GROVER BROWN

*Electrical Engineering Department
Iowa State University*

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

Preface

Copyright © 1983, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Sections 107 and 108 of the 1976 United States Copyright Act without permission of the copyright owner is unlawful. Requests for permission or further information, should be addressed to the Permissions Department, John Wiley & Sons.

Library of Congress Cataloging in Publication Data:

Brown, Robert Grover.

Introduction to random signal analysis and Kalman filtering.

Includes index.

1. Signal theory (Telecommunication) 2. Random noise theory. 3. Kalman filtering. I. Title.

TK5102.5.B696 1983 621.38'0436 82-19957

ISBN 0-471-08732-7

Printed in the United States of America

10 9 8 7 6 5 4 3

Almost all engineers get involved with instrumentation and signal processing at some time in their career. Along with signal processing comes noise—thus, it is important to understand it and know what to do about it. Noiselike signals are formally described as random processes, and a considerable amount of theory about these processes has been developed over the past 40 years. The primary emphasis here is on linear least-squares filtering. This subject has become especially important since the recursive form of the filter was introduced by R. E. Kalman in 1960. Therefore, much of the book is devoted to a topic that is now known as Kalman filtering. One of my purposes in writing this book was to bring down the level of this subject and make it available to a wider audience. Thus, wherever possible, intuitive arguments are used instead of rigorous proofs, and long derivations are avoided unless they are essential to a full understanding of the limitations of the theory. There is a strong emphasis on examples. They reinforce the theory and demonstrate its applicability to real-life engineering problems.

The study of random processes and linear filtering is no more difficult than many other subjects that are taught at the senior level in engineering. Yet, a hierarchy of prerequisite material is required and, for this reason, the subject usually gets pushed into the beginning graduate level. In particular, a working knowledge of linear system theory is needed. This includes both Laplace and Fourier transform methods and, at least, some acquaintance with state space methods. The kind of treatment given in a typical senior-level course in linear control systems is quite adequate for the level of material presented here. Noise must be described in probabilistic terms, so that at least an elementary knowledge of probability is required before proceeding on to random process theory. Since probability is a subject that engineers often miss as undergraduates, Chapter 1 fills this gap. It is a "no frills" treatment that provides the essentials needed for the remaining chapters.

Random signals and linear filtering is interdisciplinary, and this book can be used by all engineers and applied scientists, not just electrical engineers. I have taught the material to mixed groups, and I found that students who are not electrical engineers fare just as well as electrical engineers. The amount of material that can be covered in one semester, of course, depends on the background of the class. If it is first necessary to bring the group "up to speed" on probability, then it is difficult to squeeze all of the material into one semester. I

consider Chapters 8 and 9 to be optional material; they can be omitted if necessary. Also, some instructors may prefer to skip Chapter 4 on Wiener filtering and go directly to Kalman filtering. The book is organized so this may be done with little loss in continuity. In this digital age, Kalman filtering is probably the more important of the two subjects; consequently, if something has to give, it should be Wiener filtering. A minimum pedagogical objective would be to cover Chapters 1–3, 5, 6 and Section 9.1, and do it well. I consider this to be the most important material in the book from an applications viewpoint.

As you can see, the book may be used as a text in many ways, depending on the background and interests of the class. I hope the level and style will also appeal to engineers in industry as a self-study reference. Kalman filtering is an especially important topic with many potential applications, and *it is within reach of any B.S.-level engineer* with the usual background in mathematics and linear systems analysis.

I am grateful to my colleagues and students for their encouragement and helpful suggestions during the preparation of the manuscript. I also thank the office staff of the Department of Electrical Engineering for their help in typing classroom notes. I especially thank Jeanne Gehm for preparing the final manuscript.

R. GROVER BROWN

Contents

CHAPTER 1. Probability and Random Variables	1
1.1 Random Signals	1
1.2 Intuitive Notion of Probability	2
1.3 Axiomatic Probability	5
1.4 Joint and Conditional Probability	11
1.5 Independence	15
1.6 Random Variables	17
1.7 Probability Distribution and Density Functions	19
1.8 Expectation, Averages, and Characteristic Function	22
1.9 Normal or Gaussian Random Variables	26
1.10 Impulsive Probability Density Functions	31
1.11 Multiple Random Variables	32
1.12 Correlation, Covariance, and Orthogonality	39
1.13 Sum of Independent Random Variables and Tendency Toward Normal Distribution	40
1.14 Transformation of Random Variables	44
1.15 Multivariate Normal Density Function	51
1.16 Linear Transformation and General Properties of Normal Random Variables	55
1.17 Limits, Convergence, and Unbiased Estimators	58
CHAPTER 2. Mathematical Description of Random Signals	71
2.1 Concept of a Random Process	71
2.2 Probabilistic Description of a Random Process	74
2.3 Gaussian Random Process	77
2.4 Stationarity, Ergodicity, and Classification of Processes	77
2.5 Autocorrelation Function	79
2.6 Crosscorrelation Function	83
2.7 Power Spectral Density Function	84
2.8 Cross Spectral Density Function	89
2.9 White Noise	90
2.10 Gauss-Markov Process	92
2.11 Random Telegraph Wave	94
2.12 Narrowband Gaussian Process	96
2.13 Wiener or Brownian Motion Process	98
2.14 Pseudorandom Signals	100

2.15	Determination of Autocorrelation and Spectral Density Functions from Experimental Data	103	6.5	Power System Relaying Application	229
2.16	Sampling Theorem	109	6.6	Divergence Problems	233
2.17	Discrete Fourier Transform and Fast Fourier Transform	110	6.7	Off-Line System Error Analysis	239
CHAPTER 3. Response of Linear Systems to Random Inputs		122	6.8	Relationship to Deterministic Least Squares	242
3.1	Introduction: The Analysis Problem	122	6.9	Telephone Load Forecasting Example	245
3.2	Stationary (Steady-State) Analysis	123	6.10	Deterministic Inputs	248
3.3	Integral Tables for Computing Mean-Square Value	126	CHAPTER 7. The Continuous Kalman Filter		253
3.4	Pure White Noise and Bandlimited Systems	127	7.1	Transition from the Discrete to Continuous Filter Equations	253
3.5	Noise Equivalent Bandwidth	128	7.2	Solution of the Matrix Riccati Equation	257
3.6	Shaping Filter	130	7.3	Correlated Measurement and Process Noise	259
3.7	Nonstationary (Transient) Analysis—Initial Condition Response	131	7.4	Colored Measurement Noise	262
3.8	Nonstationary (Transient) Analysis—Forced Response	133	7.5	Suboptimal Error Analysis	266
3.9	Summary	136	7.6	Relationship Between Wiener and Kalman Filters	267
CHAPTER 4. Wiener Filtering		147	CHAPTER 8. Discrete Smoothing and Prediction		273
4.1	Perspective	147	8.1	Prediction	273
4.2	Optimization with Respect to a Parameter	148	8.2	Classification of Smoothing Problems	274
4.3	The Stationary Optimization Problem—Weighting Function Approach	151	8.3	Discrete Fixed-Interval Smoothing	275
4.4	The Nonstationary Problem	160	8.4	Discrete Fixed-Point Smoothing	280
4.5	Orthogonality	164	8.5	Fixed-Lag Smoothing	283
4.6	Complementary Filter	165	8.6	Forward-Backward Filter Approach to Smoothing	285
4.7	Multiple-Input Wiener Filter Problem	169	CHAPTER 9. Additional Topics on Kalman Filtering		296
4.8	Weighting Function Approach to the Sampled-Data Filter	171	9.1	Linearization of Nonlinear Process Dynamics	296
4.9	Perspective	173	9.2	Adaptive (Self-learning) Kalman Filter	302
CHAPTER 5. The Discrete Kalman Filter		181	9.3	Square Root Filtering	307
5.1	A Simple Recursive Example	181	9.4	Conversion of Continuous Measurements to Discrete Form—The Delayed State Filter Algorithm	312
5.2	Vector Description of a Continuous Random Process	183	9.5	Stochastic Linear Regulator Problem and the Separation Theorem	320
5.3	Discrete-Time Model	188	APPENDIX Laplace and Fourier Transforms		327
5.4	The Discrete Kalman Filter	195	A.1	The One-Sided Laplace Transform	327
5.5	Examples	200	A.2	The Fourier Transform	330
5.6	The Gaussian Process and Conditional Density Viewpoint	204	A.3	Two-Sided Laplace Transform	332
CHAPTER 6. Applications and Additional Topics on Discrete Kalman Filtering		213	INDEX		341
6.1	Augmenting the State Vector	213			
6.2	Alternative Form of the Discrete Kalman Filter	217			
6.3	Sequential Processing of the Measurement Data	220			
6.4	Aided Inertial Navigation Systems	222			

INTRODUCTION TO

**RANDOM SIGNAL
ANALYSIS AND
KALMAN FILTERING**

Probability and Random Variables

1.1 RANDOM SIGNALS

Nearly everyone has some notion of random or noiselike signals. One has only to tune an ordinary AM radio away from a station, turn up the volume, and the result is static, or noise. If one were to look at a strip-chart recording of such a signal, it would appear to wander on aimlessly with no apparent order in its amplitude pattern as shown in Fig. 1.1. Signals of this type cannot be described with explicit mathematical functions such as sine waves, step functions, and the like. Their description must be put in probabilistic terms. Early investigators recognized that random signals could be described loosely in terms of their spectral content, but a rigorous mathematical description of such signals was not formulated until the 1940s, most notably with the work of Wiener and Rice (1,2).

Noise is usually unwanted. The additive noise in the radio signal disturbs our enjoyment of the music or interferes with our understanding of the spoken word; noise in an electronic navigation system induces position errors that can be disastrous in critical situations; noise in a digital data transmission system can cause bit errors with obvious undesirable consequences; and on and on. Any noise that corrupts the desired signal is bad; it is just a question of how bad! Even after the designer has done his best to eliminate all the obvious noise-producing mechanisms, there always seems to be some noise left over that must be suppressed with more subtle means, such as filtering. To do so effectively, one must understand noise in quantitative terms.

Probability plays a key role in the description of noiselike signals. Our treatment of this subject must necessarily be brief and directed toward the specific needs of subsequent chapters. The scope is thus limited in this regard. We make no apology for this, because many fine books have been written on probability in the broader sense. Our main objective here is the study of random signals and optimal filtering, and we wish to move on to this area as quickly as possible. First, though, we must at least review the bare essentials of probability theory.



Figure 1.1. Typical noise waveform.

1.2 INTUITIVE NOTION OF PROBABILITY

Most engineering and science students have had some acquaintance with the intuitive concepts of probability. Typically, with the intuitive approach we first consider all possible outcomes of a chance experiment as being equally likely, and then the probability of a particular event, say event A , is defined as

$$P(A) = \frac{\text{Possible outcomes favoring event } A}{\text{Total possible outcomes}} \quad (1.2.1)$$

where we read $P(A)$ as "probability of event A ." This concept is then expanded to include the relative-frequency-of-occurrence or statistical viewpoint of probability. With the relative-frequency concept, we imagine a large number of trials of some chance experiment and then define probability as the relative frequency of occurrence of the event in question. Considerations such as what is meant by "large" and the existence of limits are normally avoided in elementary treatments. This is for good reason. The idea of limit in a probabilistic sense is subtle.

Although the older intuitive notions of probability have limitations, they still play an important role in probability theory. The ratio-of-possible-events concept is a useful problem-solving tool in many instances. The relative-frequency concept is especially helpful in visualizing the statistical significance of the results of probability calculations. That is, it provides the necessary tie between the theory and the physical situation. Two examples that illustrate the usefulness of these intuitive notions of probability should now prove useful.

Example 1.1 In straight poker, each player is dealt 5 cards face down from a deck of 52 playing cards. We pose two questions:

- What is the probability of being dealt four of a kind, that is, four aces, four kings, and so forth?
- What is the probability of being dealt a straight flush, that is, a continuous sequence of five cards in any suit? ■

SOLUTION TO QUESTION (a) This problem is relatively complicated if you think in terms of the sequence of chance events that can take place when the cards are dealt one at a time. Yet the problem is relatively easy when viewed in terms

of the ratio of favorable to total number of outcomes. These are easily counted in this case. There are only 48 possible hands containing 4 aces; another 48 containing 4 kings; etc. Thus, there are $13 \cdot 48$ possible four-of-a-kind hands. The total number of possible poker hands of any kind is obtained from the combination formula for "52 things taken 5 at a time" (3). This is given by the binomial coefficient

$$\binom{52}{5} = \frac{52!}{5!(52-5)!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 2,598,960 \quad (1.2.2)$$

Therefore, the probability of being dealt four of a kind is

$$P(\text{Four of a kind}) = \frac{13 \cdot 48}{2,598,960} = \frac{624}{2,598,960} \approx .00024 \quad (1.2.3)$$

SOLUTION TO QUESTION (b) Again, the direct itemization of favorable events is the simplest approach. The possible sequences in each of four suits are: AKQJ10, KQJ109, . . . , 5432A. (*Note.* we allow the ace to be counted either high or low.) Thus, there are 10 possible straight flushes in each suit (including the royal flush of the suit) giving a total of 40 possible straight flushes. The probability of a straight flush is, then,

$$P(\text{Straight flush}) = \frac{40}{2,598,960} \approx .000015 \quad (1.2.4)$$

We note in passing that in poker a straight flush wins over four of a kind; and, rightly so, since it is the rarer of the two hands.

Example 1.2. Craps is a popular gambling game played in casinos throughout the world (11). The player rolls two dice and plays against the house (i.e., the casino). If the first roll is 7 or 11, the player wins immediately; if it is 2, 3 or 12, the player loses immediately. If the first roll results in 4, 5, 6, 8, 9, or 10, the player continues to roll until either the same number appears, which constitutes a win, or a 7 appears, which results in the player losing. What is the player's probability of winning when throwing the dice?

This example was chosen to illustrate the shortcoming of the direct count-the-outcomes approach. In this case, one cannot enumerate all the possible outcomes. For example, if the player's first roll is a 4, the play continues until another outcome is reached. Presumably, the rolling could continue on ad infinitum without a 4 or 7 appearing, which is what is required to terminate the game. Thus, the direct enumeration approach fails in this situation. On the other hand, the relative-frequency-of-occurrence approach works quite well. Table 1.1 shows the relative frequency of occurrence of the various numbers on the first roll. The numbers in the column labeled "probability" were obtained by enumerating the 36 possible outcomes and allotting $\frac{1}{36}$ for each outcome that

Table 1.1 Probabilities in Craps

Number of First Throw	Probability	Result of First Throw	Subsequent Probabilities and Results	Relative Frequency of Winning with Various First Throws
2	$\frac{1}{36}$	Lose		0
3	$\frac{2}{36}$	Lose		0
4	$\frac{3}{36}$	Continue	$P(4 \text{ before } 7) = \frac{1}{2} \text{ (win)}$ $P(7 \text{ before } 4) = \frac{2}{3} \text{ (lose)}$	$\frac{3}{36} \cdot \frac{1}{2}$
5	$\frac{4}{36}$	Continue	$P(5 \text{ before } 7) = \frac{2}{3} \text{ (win)}$ $P(7 \text{ before } 5) = \frac{2}{3} \text{ (lose)}$	$\frac{4}{36} \cdot \frac{2}{3}$
6	$\frac{5}{36}$	Continue	$P(6 \text{ before } 7) = \frac{1}{11} \text{ (win)}$ $P(7 \text{ before } 6) = \frac{6}{11} \text{ (lose)}$	$\frac{5}{36} \cdot \frac{1}{11}$
7	$\frac{6}{36}$	Win		$\frac{6}{36}$
8	$\frac{5}{36}$	Continue	$P(8 \text{ before } 7) = \frac{6}{11} \text{ (win)}$ $P(7 \text{ before } 8) = \frac{5}{11} \text{ (lose)}$	$\frac{5}{36} \cdot \frac{6}{11}$
9	$\frac{4}{36}$	Continue	$P(9 \text{ before } 7) = \frac{2}{3} \text{ (win)}$ $P(7 \text{ before } 9) = \frac{2}{3} \text{ (lose)}$	$\frac{4}{36} \cdot \frac{2}{3}$
10	$\frac{3}{36}$	Continue	$P(10 \text{ before } 7) = \frac{1}{3} \text{ (win)}$ $P(7 \text{ before } 10) = \frac{2}{3} \text{ (lose)}$	$\frac{3}{36} \cdot \frac{1}{3}$
11	$\frac{2}{36}$	Win		$\frac{2}{36}$
12	$\frac{1}{36}$	Lose		0

$$\text{Total probability of winning} = \frac{24}{36} = .4929$$

yields a sum corresponding to the number in the first column. For example, a 4 may be obtained with the combinations (1,3), (2,2), or (1,3). For the cases where the game continues after the first throw, the subsequent probabilities were obtained simply by observing the *relative* frequency of occurrence of the numbers involved. For example, a 7 is twice as likely as a 4. Thus, the relative frequency of rolling a 7 before a 4 should be twice that of "4 before 7," and the respective probabilities are $\frac{2}{3}$ and $\frac{1}{2}$. The total probability of winning with a 4 on the first throw was reasoned as follows. A 4 only appears on the first roll $\frac{3}{36}$ of

the time; and, of this fraction, only $\frac{1}{2}$ of the time will this result in an ultimate win. Thus, the relative frequency of winning via this route is the product of $\frac{3}{36} \cdot \frac{1}{2}$. Admittedly, this line of reasoning is quite intuitive, but that is the very nature of the relative-frequency-of-occurrence approach to probability.

For the benefit of those who like to gamble, it should be noted that craps is a very close game. The edge in favor of the house is only about $1\frac{1}{2}$ percent. (Also see Problem 1.7.) ■

1.3 AXIOMATIC PROBABILITY

It should be apparent that the intuitive concepts of probability have their limitations. The ratio-of-outcomes approach requires the equal-lielihood assumption for all outcomes. This may fit many situations, but often we wish to consider "unfair" chance situations as well as "fair" ones. Also, as demonstrated in Example 1.2, there are many problems for which all possible outcomes simply cannot be enumerated. The relative-frequency approach is intuitive by its very nature. Intuition should never be ignored; but, on the other hand, it can lead one astray in complex situations. For these reasons, the axiomatic formulation of probability theory is now almost universally favored among both applied and theoretical scholars in this area. As we would expect, axiomatic probability is compatible with the older, more heuristic probability theory.

Axiomatic probability begins with the concept of a *sample space*. We first imagine a conceptual chance experiment. The sample space is the set of all possible *outcomes* of this experiment. The individual outcomes are called *elements* or *points* in the sample space. We denote the sample space as S and its set of elements as $\{s_1, s_2, s_3, \dots\}$. The number of points in the sample space may be finite, countably infinite, or simply infinite, depending on the experiment under consideration. A few examples of sample spaces should be helpful at this point.

Example 1.3 *The experiment:* Make a single draw from a deck of 52 playing cards. Since there are 52 possible outcomes, the sample space contains 52 discrete points. If we wished, we could enumerate them as Ace of Clubs, King of Clubs, Queen of Clubs, and so forth. Note that the points of the sample space in this case are "things," not numbers. ■

Example 1.4 *The experiment:* Two fair dice are thrown and the number of dots on the top of each is observed. There are 36 discrete outcomes that can be enumerated as (1,1), (1,2), (1,3), . . . , (6,5), (6,6). The first number in paren-

theses identifies the number of dots on die 1 and the second is the number on die 2. Thus, 36 distinct 2-tuples describe the possible outcomes, and our sample space contains 36 points or elements. Note that the points in this sample space retain the identity of each individual die and the number of dots shown on its top face. ■

Example 1.5 *The experiment:* Two fair dice are thrown and the sum of the number of dots is observed. In this experiment, we do not wish to retain the identity of the numbers on each die; only the sum is of interest. Therefore, it would be perfectly proper to say the possible outcomes of the experiment are $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Thus, the sample space would contain 11 discrete elements. From this and the preceding example, it can be seen that we have some discretion in how we define the sample space corresponding to a certain experiment. It depends to some extent on what we wish to observe. If certain details of the experiment are not of interest, they often may be suppressed with some resultant simplification. However, once we agree on what items are to be grouped together and called *outcomes*, the sample space must include all the defined outcomes; and, similarly, the result of an experiment must always yield one of the defined outcomes. ■

Example 1.6 *The experiment:* A dart is thrown at a target and the location of the hit is observed. In this experiment we imagine the random mechanisms affecting the throw are such that we get a continuous spread of data centered around the bull's-eye when the experiment is repeated over and over. In this case, even if we bound the hit locations within a certain region determined by reasonableness, we still cannot enumerate all possible hit locations. Thus, we have an infinite number of points in our sample space in this example. Even though we cannot enumerate the points one by one, they are, of course, identifiable in terms of either rectangular or polar coordinates. ■

It should be noted that elements of a sample space must always be *mutually exclusive* or *disjoint*. On a given trial, the occurrence of one excludes the occurrence of another. There is no overlap of points in a sample space.

In axiomatic probability, the term event has special meaning and should not be used interchangably with outcome. An *event* is a special subset of the sample space S . We usually wish to consider various events defined on a sample space, and they will be denoted with uppercase letters such as A, B, C, \dots , or perhaps A_1, A_2, \dots , etc. Also, we will have occasion to consider the set of operations of union, intersection, and complement of our defined events. Thus, we must be careful in our definition of events to make the set sufficiently complete such that these set operations also yield properly defined events. In discrete problems, this can always be done by defining the set of events under consideration to be all possible subsets of the sample space S . We will tacitly

assume that the null set is a subset of every set, and that every set is a subset of itself.

One other comment about events is in order before proceeding to the basic axioms of probability. The event A is said to occur if *any* point in A occurs.

The three axioms of probability may now be stated. Let S be the sample space and A be any event defined on the sample space. The first two axioms are

$$\text{Axiom 1: } P(A) \geq 0 \quad (1.3.1)$$

$$\text{Axiom 2: } P(S) = 1 \quad (1.3.2)$$

Now, let A_1, A_2, A_3, \dots be mutually exclusive (disjoint) events defined on S . The sequence may be finite or countably infinite. The third axiom is then

$$\begin{aligned} \text{Axiom 3: } & P(A_1 \cup A_2 \cup A_3 \cup \dots) \\ & = P(A_1) + P(A_2) + P(A_3) + \dots \end{aligned} \quad (1.3.3)$$

Axiom 1 simply says that the probability of an event cannot be negative. This certainly conforms to the relative-frequency-of-occurrence concept of probability. Axiom 2 says that the event S , which includes all possible outcomes, must have a probability of unity. It is sometimes called the certain event. The first two axioms are obviously necessary if axiomatic probability is to be compatible with the older relative-frequency probability theory. The third axiom is not quite so obvious, perhaps, and it simply must be assumed. In words, it says that when we have nonoverlapping (disjoint) events, the probability of the union of these events is the sum of the probabilities of the individual events. If this were not so, one could easily think of counterexamples that would not be compatible with the relative-frequency concept. This would be most undesirable.

We now recapitulate. There are three essential ingredients in the formal approach to probability. First, a sample space must be defined that includes all possible outcomes of our conceptual experiment. We have some discretion in what we call outcomes, but caution is in order here. The outcomes must be disjoint and all-inclusive such that $P(S) = 1$. Second, we must carefully define a set of events on the sample space, and the set must be closed such that the operations of union, intersection, and complement also yield events in the set. Finally, we must assign probabilities to all events in accordance with the basic axioms of probability. In physical problems, this assignment is chosen to be compatible with what we feel to be reasonable in terms of relative frequency of occurrence of the events. If the sample space S contains a finite number of elements, the probability assignment is usually made directly on the elements of S . They are, of course, elementary events themselves. This, along with Axiom 3, then indirectly assigns a probability to all other events defined on the sample space. However, if the sample space consists of an infinite "smear" of points, the probability assignment must be made on events and not on points in the sample space. This will be illustrated later in Example 1.8.

Once we have specified the sample space, the set of events, and the probabilities associated with the events, we have what is known as a *probability space*. This provides the theoretical structure for the formal solution of a wide variety of probability problems.

Example 1.7 Consider a single throw of two dice, and let us say we are only interested in the sum of the dots that appear on the top faces. This chance situation fits many games that are played with dice. In this case, we will define our sample space to be

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

and it is seen to contain 11 discrete points. Next, we define the set of possible events to be all subsets of S , including the null set and S itself. Note that the elements of S are elementary events, and they are disjoint, as they should be. Also, $P(S) = 1$. Finally, we need to assign probabilities to the events. This could be done arbitrarily (within the constraints imposed by the axioms of probability), but in this case we want the results of our formal analysis to coincide with the relative-frequency approach. Therefore, we will assign probabilities to the elements in accordance with Table 1.2, which, in turn, indirectly specifies probabilities for all other events defined on S . We now have a properly defined probability space, and we can pose a variety of questions relative to the single throw of two dice.

Suppose we ask: What is the probability of throwing either a 7 or an 11? From Axiom 3, and noting that "7 or 11" is the equivalent of saying " $7 \cup 11$," we have

$$P(7 \text{ or } 11) = \frac{6}{36} + \frac{2}{36} = \frac{2}{9} \quad (1.3.4)$$

Table 1.2 Probabilities for Two-Dice Example

Sum of Two Dice	Assigned Probability
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$
5	$\frac{4}{36}$
6	$\frac{5}{36}$
7	$\frac{6}{36}$
8	$\frac{5}{36}$
9	$\frac{4}{36}$
10	$\frac{3}{36}$
11	$\frac{2}{36}$
12	$\frac{1}{36}$

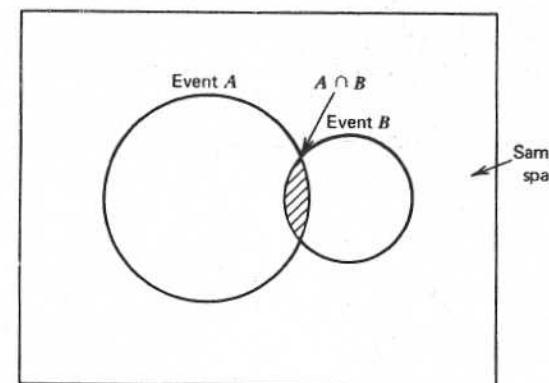


Figure 1.2. Venn diagram for two events A and B .

Next, suppose we ask: What is the probability of not throwing 2, 3, or 12? This calls for the complement of event "2 or 3 or 12" which is the set $\{4, 5, 6, 7, 8, 9, 10, 11\}$. Recall that we say the event occurs if *any* element in the set occurs. Therefore, again using Axiom 3, we have

$$\begin{aligned} P(\text{Not throwing } 2, 3, \text{ or } 12) &= \frac{3 + 4 + 5 + 6 + 5 + 4 + 3 + 2}{36} \\ &= \frac{8}{9} \end{aligned} \quad (1.3.5)$$

Suppose we now pose the further question: What is the probability that two "4s" are thrown? In our definition of the sample space, we suppressed the identity of the individual dice, so this simply is not a proper question for the probability space, as defined. This example will be continued, but first we digress for a moment to consider intersection of events. ■

In addition to the set operations of union and complementation, the operation of intersection is also useful in probability theory. The intersection of two events A and B is the event containing points that are common to both A and B . This is illustrated in Fig. 1.2 with what is sometimes called a Venn diagram. The points lying within the heavy contour comprise the union of A and B , denoted as $A \cup B$ or " A or B ." The points within the shaded region are the event " A intersection B ," which is denoted $A \cap B$, or sometimes just " A and B ."^{*} The following relationship should be apparent just from the geometry of

* In many references, the notation for " A union B " is " $A + B$," and " A intersection B " is shortened to just " AB ." We will be proceeding to the study of random variables shortly, and then the chance occurrences will be related to real numbers, not "things." Thus, in order to avoid confusion, we will stay with the more cumbersome notation of \cup and \cap for set operations, and reserve $X + Y$ and XY to mean the usual arithmetic operations on real variables.

the Venn diagram:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (1.3.6)$$

The subtractive term in Eq. (1.3.6) is present because the probabilities in the overlapping region have been counted twice in the summation of $P(A)$ and $P(B)$.

The probability $P(A \cap B)$ is known as the *joint* probability of A and B and will be discussed further in Section 1.4. We digress for the moment, though, and look at two examples.

Example 1.7 (continued) We return to the two-dice example. Suppose we define event A as throwing a 4, 5, 6, or 7. Event B will be defined as throwing a 7, 8, 9, 10, or 11. We now pose two questions:

- (a) What is the probability of event "A and B" (i.e., $A \cap B$)?
- (b) What is the probability of event "A or B" (i.e., $A \cup B$)?

The answer to (a) is found by looking for the elements of the sample space that are common to both events A and B . Since the number 7 is the only common element,

$$P(A \cap B) = P(7) = \frac{1}{6}$$

The answer to (b) can be found either by itemizing the elements in $A \cup B$ or from Eq. (1.3.6). Using Eq. (1.3.6) leads to

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{18}{36} + \frac{20}{36} - \frac{6}{36} \\ &= \frac{8}{9} \end{aligned}$$

This is easily verified using the direct itemization method. ■

Example 1.8 Let us reconsider the dart-throwing experiment. We might consider a simple game where we draw a circle of radius R on the wall. If the player's dart lands on or within the circle, the player wins; if the dart lands outside R , the player loses. We pose the simple question: What is the probability that the player wins on the single throw of a dart?

In this example, the sample space is all the points on the wall. (We assume that the player can at least hit the wall.) Thus, there are an infinite number of points in our sample space S .

In this game, we only define two events on the sample space—event W , player wins; and event L , player loses. This is a legitimate set of events because the set operations of union, intersection, and complement yield defined events of the set.

The assignment of probabilities in this case must be made directly on the two events rather than on the sample space, because the probability of hitting *exactly* any particular point on the wall is zero and this tells us nothing about events. To add a note of realism, we might speculate that we are devising a simple gambling game for the house, one that the patrons will enjoy and that will also be profitable to the house. Our observations indicate that the typical player throws about 10 percent of his or her darts within a circle of radius R_1 . In addition, the hits are more or less uniformly spaced within the circle. It would then be appropriate to assign probabilities to the two events as

$$P(\text{Hit lies on or within } R_1) = .1$$

$$P(\text{Hit lies outside } R_1) = .9$$

If the establishment were to offer 5 to 1 odds, this game might well please the players and at the same time produce revenue for the house.

Before leaving this example, it should be mentioned that the radius R may be treated as a parameter, and hence we have the structure for looking at a whole family of problems, not just a single one. To add variety to the game, the proprietor might occasionally wish to decrease the diameter of the circle and increase the odds. Clearly, if the hits within the circle are nearly uniformly distributed, reducing the area by a factor of 2 will reduce the relative frequency of occurrence by a similar factor. Yet, the corresponding reduction in radius is only a factor of $\sqrt{2}$, which should make the game "look" attractive. The appropriate probability assignments to this case would be .05 for "win" and .95 for "lose." Presumably, the house could offer 10 to 1 odds and still net the same average return as received on the 5-to-1 game. ■

1.4 JOINT AND CONDITIONAL PROBABILITY

In complex situations it is often desirable to arrange the elements of the sample space in arrays. This provides an orderly way of grouping related points in the space and is especially useful when considering successive trials of similar experiments. Consider the following example.

Example 1.9 *The experiment:* Draw a card from a deck of 52 playing cards. Then replace the card, reshuffle the deck, and draw a card a second time. This is known as sampling with replacement. The sample space for this experiment, when viewed as a whole, consists of all possible pairs of cards for the two draws. This amounts to $(52)^2$ or 2704 elements—clearly an unwieldy number of elements to keep track of without some systematic way of ordering things. Yet, in spite of its size, the sample space is quite manageable when viewed as an array as shown in Table 1.3. This is a two-dimensional array, but it should be obvious that the concept is easily extended to higher-order situations—concep-

Table 1.3 Joint Probabilities for Two Draws with Replacement

First Draw \ Second Draw	Ace of Spades	King of Spades	...	Deuce of Clubs
Ace of spades	$\frac{1}{2704}$	$\frac{1}{2704}$	$\frac{1}{2704}$	$\frac{1}{2704}$
King of spades	$\frac{1}{2704}$	$\frac{1}{2704}$	$\frac{1}{2704}$	$\frac{1}{2704}$
...
(52 elements)		(Total of 2704 entries)		
Deuce of clubs	$\frac{1}{2704}$	$\frac{1}{2704}$	$\frac{1}{2704}$	$\frac{1}{2704}$

tually at least. Had we specified three draws rather than two, we would have a three-dimensional array, and so forth. Thus the idea of an n -dimensional array associated with n trials of an experiment is an important concept. Note that summing out the numbers along any particular row yields the probability of drawing that particular card on the first draw irrespective of the result of the second draw. This leads to the idea of marginal probability, which will now be considered in a more general setting. ■

Let A and B loosely refer to “first” and “second” chance experiments. Time is not of the essence, so the experiments may be either successive or simultaneous as the situation dictates. Let A_1, A_2, \dots, A_m denote disjoint events associated with experiment A and, similarly, B_1, B_2, \dots, B_n denote disjoint events for experiment B . This leads to the joint-probability array shown in Table 1.4. Note first that n does not have to equal m , and therefore the array is not necessarily square. Also, the so-called *marginal probabilities* are shown in Table 1.4 to the right and bottom of the joint-probability array. These are the result of summing out rows or columns, as the case may be, and the term marginal arose because these probabilities are often written in the margins outside the $m \times n$ array. Clearly, summing out horizontally yields the probability of a particular event in the experiment A , irrespective of results of experiment B . Similarly, summing columns yields $P(B_1), P(B_2)$, and so forth. Also, we tacitly assume that events A_1, A_2, \dots, A_m and B_1, B_2, \dots, B_n are all-inclusive as well as disjoint, so that the sum of the marginal probabilities, either vertically or horizontally, is unity.

Table 1.4 also contains information about the relative frequency of occurrence of various events in one set, given a particular event in the other set. For example, look at column 2, which lists $P(A_1 \cap B_2), P(A_2 \cap B_2), \dots, P(A_m \cap B_2)$. Since no other entries in the table involve B_2 , this list of numbers gives the relative distribution of events A_1, A_2, \dots, A_m , given B_2 has occurred. How-

Table 1.4 Array of Joint and Marginal Probabilities

$A \backslash B$	Event B_1	Event B_2	...	Event B_n	Marginal Probabilities
Event A_1	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$...	$P(A_1 \cap B_n)$	$P(A_1)$
Event A_2	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$...	$P(A_2 \cap B_n)$	$P(A_2)$
...
Event A_m	$P(A_m \cap B_1)$	$P(A_m \cap B_2)$...	$P(A_m \cap B_n)$	$P(A_m)$
Marginal Probabilities	$P(B_1)$	$P(B_2)$...	$P(B_n)$	Sum = 1

ever, the set of numbers appearing in column 2 is not a legitimate probability distribution because the sum is $P(B_2)$, not unity. So, imagine “renormalizing” all the entries in the column by dividing by $P(B_2)$. The new set of numbers is then $P(A_1 \cap B_2)/P(B_2), P(A_2 \cap B_2)/P(B_2), \dots, P(A_m \cap B_2)/P(B_2)$, the sum is unity, and the relative distribution corresponds to the relative frequency of occurrence of A_1, A_2, \dots, A_m , given B_2 . This heuristic reasoning leads us to the formal definition of conditional probability.

The *conditional probability* of A_i given B_j is defined as

$$P(A_i|B_j) = \frac{P(A_i \cap B_j)}{P(B_j)} \quad (1.4.1)$$

Similarly, the conditional probability of B_j given A_i is

$$P(B_j|A_i) = \frac{P(B_j \cap A_i)}{P(A_i)} \quad (1.4.2)$$

It is tacitly assumed in the above equations that $P(B_j)$ and $P(A_i)$ are not zero. Otherwise, conditional probability is not defined. It should also be emphasized that conditional probability is a *defined* concept and is not derived from other concepts. The discussion leading up to Eqs. (1.4.1) and (1.4.2) was presented to give an intuitive rationale for the definition and was not intended as a proof.

A useful relationship is obtained when Eqs. (1.4.1) and (1.4.2) are combined. Each equation may be solved for the probability of A_i intersection B_j and the results equated. This leads to *Bayes rule* (or *Bayes theorem*):

$$P(A_i|B_j) = \frac{P(B_j|A_i) P(A_i)}{P(B_j)} \quad (1.4.3)$$

This relationship is useful in reversing the conditioning of events. Note that the joint probability array $P(A_i \cap B_j)$ contains all the necessary information for computing *all* marginal and conditional probabilities. Conversely, if you know

Table 1.5 Joint Probability Table for Four-Ball Urn Example

		Second Draw			
		Red 1	Red 2	Black 1	Black 2
First Draw	Red 1	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
	Red 2	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
Black 1	Red 1	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
	Black 2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0

$P(A_i|B_j)$ and $P(B_j)$ [or $P(B_j|A_i)$ and $P(A_i)$], there is sufficient information to find $P(A_i \cap B_j)$.

Example 1.10 For variety, we now consider an urn problem. The urn contains two red and two black balls. Two balls are drawn sequentially from the urn without replacement.

- (a) What is the array of joint probabilities for the first and second draws?
- (b) What is the conditional probability that the second draw is red, given the first draw is red?

To obtain the joint probability table, we first define a sample space consisting of all possible outcomes, including the identity of the individual balls. The four balls will be referred to as Red 1, Red 2, Black 1, and Black 2. The joint probability array for the first and second draws is given in Table 1.5. Note the effect of the “without replacement” statement. It gives rise to zeros along the major diagonal, because drawing Red 1 on the first draw precludes Red 1 being drawn again on the second draw, and so forth. In effect, there are really only 12 nontrivial outcomes for this experiment, and we assume they are all equally likely.

In the original problem, there was no mention of retaining the individuality of the two red and two black balls—only “red” and “black” were specified. Therefore, we can consolidate outcomes in accordance with the partitioning shown by dashed lines in Table 1.5. This leads to the two-by-two array shown in Table 1.6. This, then, is the answer to part (a).

For the conditional probability we will use the basic definition given by Eq. (1.4.2). Writing this out explicitly for the conditional situation posed in question (b), we have

$$P(\text{Second draw red}|\text{First draw red}) = \frac{P(\text{First draw red and second draw red})}{P(\text{First draw red})} \quad (1.4.4)$$

The numerator of Eq. (1.4.4) is the upper-left entry in Table 1.6. The denomina-

Table 1.6 Joint Probability Table Reduced to Two-by-Two Array

		Second Draw	
		Red	Black
First Draw	Red	$\frac{1}{2}$	$\frac{1}{2}$
	Black	$\frac{1}{2}$	$\frac{1}{2}$

tor is the marginal probability obtained by summing the elements of the first row in Table 1.6. This yields

$$P(\text{Second draw red}|\text{First draw red}) = \frac{\frac{1}{2}}{\frac{1}{2}} = \frac{1}{2} \quad (1.4.5)$$

This is the solution to part (b). Note this checks with the result one would obtain by considering the three balls that remain in the urn after a red one is withdrawn. ■

1.5 INDEPENDENCE

In qualitative terms, two events are said to be independent if the occurrence of one does not affect the likelihood of the other. If we toss two coins simultaneously, we would not expect the outcome of one of the coins to affect the other. Similarly, if we draw a card from a deck of 52 playing cards, then replace it, reshuffle, and draw a second time, we would not expect the second outcome to be affected by the first. However, if we draw the second card without replacing the first, it is a much different matter. For example, the probability of drawing an ace on the second draw with replacement is $\frac{4}{52}$. However, if we draw an ace on the first draw and do not replace it, the probability of getting an ace on the second draw is only $\frac{3}{51}$. In the “without replacement” experiment, the outcome of the first draw certainly affects the chances on the second draw, so the two events are not independent.

Formally, events A and B are said to be *independent* if

$$P(A \cap B) = P(A)P(B) \quad (1.5.1)$$

Also, it should be evident from Eq. (1.5.1) and the defining equations for conditional probability, Eqs. (1.4.1) and (1.4.2), that if A and B are independent

$$\begin{aligned} P(A|B) &= P(A) && \text{For } A \text{ and } B \\ P(B|A) &= P(B) && \text{independent only} \end{aligned} \quad (1.5.2)$$

Table 1.7 Joint and Marginal Probabilities for Toss of Two Coins

		Second Coin	
		Heads	Tails
First Coin	Heads	$\frac{1}{4}$	$\frac{1}{4}$
	Tails	$\frac{1}{4}$	$\frac{1}{4}$
		$\frac{1}{2}$ (Prob. second coin is heads)	$\frac{1}{2}$ (Prob. second coin is tails)
		$\frac{1}{2}$ (Prob. first coin is heads)	$\frac{1}{2}$ (Prob. first coin is tails)

We might also note that the defining equation for independence, Eq. (1.5.1), usually provides the simplest test for independence. This is illustrated with two examples.

Example 1.11 The joint probability array for the simultaneous toss of two coins is given in Table 1.7. The marginal probabilities are also shown in the “margins” with their significance stated in words in parentheses. Note that each of the four joint probabilities of the array may be written as the product of their respective marginal probabilities. Thus, all events are independent in this case.

Example 1.12 Let us reconsider the urn experiment described in Example 1.10, Section 1.4. Recall that the two balls were withdrawn sequentially *without replacement*, and that this led to the joint probability array shown in Table 1.8. The marginal probabilities are also included, just as in the previous example. However, in this case none of the joint probabilities can be written as the product of the respective marginal probabilities. Thus, all event pairs are dependent. ■

Table 1.8 Joint and Marginal Probabilities for Urn Example

		Second Draw	
		Red	Black
First Draw	Red	$\frac{1}{3}$	$\frac{2}{3}$
	Black	$\frac{2}{3}$	$\frac{1}{3}$
		$\frac{1}{2}$	$\frac{1}{2}$

To recapitulate, we say events A and B are independent if their joint probability can be written as the product of the individual total probabilities, $P(A)$ and $P(B)$. Otherwise, they are said to be dependent.

1.6 RANDOM VARIABLES

In the study of noiselike signals, we are nearly always dealing with physical quantities such as voltage, torque, distance, and so forth, which can be measured in physical units. In these cases, the chance occurrences are related to real numbers, not just “things” like heads or tails, black balls or red balls, etc. This brings us to the notion of a random variable. Let us say we have a conceptual experiment for which we have defined a suitable sample space, an appropriate set of events, and a probability assignment for the set of events. A *random variable* is simply a function that maps every point in the sample space (things) on to the real line (numbers). A simple example of this mapping is shown in Fig. 1.3. Note that each face of the die is embossed with a pattern of dots, not a number. The assignment of numbers is our own doing and could be most anything. In Fig. 1.3 we just happened to choose the most common numerical assignment, namely the sum of the number of dots, but this was not necessary.

Presumably, in our probability space, probabilities have been assigned to the events in the sample space in accordance with the basic axioms of probability. Associated with each event in the original sample space (things) there will be a corresponding event in the random-variable space (numbers). These will

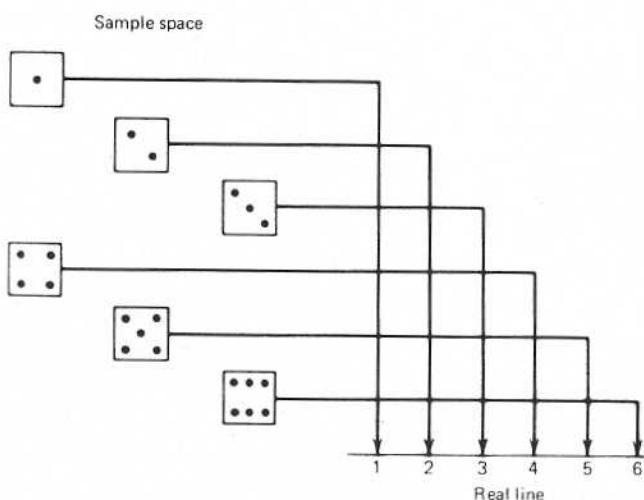


Figure 1.3. Mapping for the throw of one die.

be called *equivalent events*, and it is only natural that we should assign probabilities to the random-variable events in the same manner as for the original sample-space events. Stated formally, we have

$$\begin{aligned} P(\text{equivalent event on the real line}) \\ = P(\text{corresponding event in the original sample space}) \end{aligned} \quad (1.6.1)$$

Two more examples will illustrate the concept of a random variable further.

Example 1.13 The mapping that defines a random variable must fit the chance situation at hand if it is to be useful. Sometimes this leads to unusual but perfectly legitimate functional relationships. For example, in the game of pitch, a portion of the scoring is done by summing the card values of the cards each player takes in tricks during the course of play. The card values, by arbitrary rules of the game, are as follows:

Card of Any Suit	Card Value
2 through 9	0
10	10
Jack	1
Queen	2
King	3
Ace	4

Thus, in exploring your chances relative to this aspect of the game, it would be appropriate to map the 52 points in the sample space (for a single card) into real

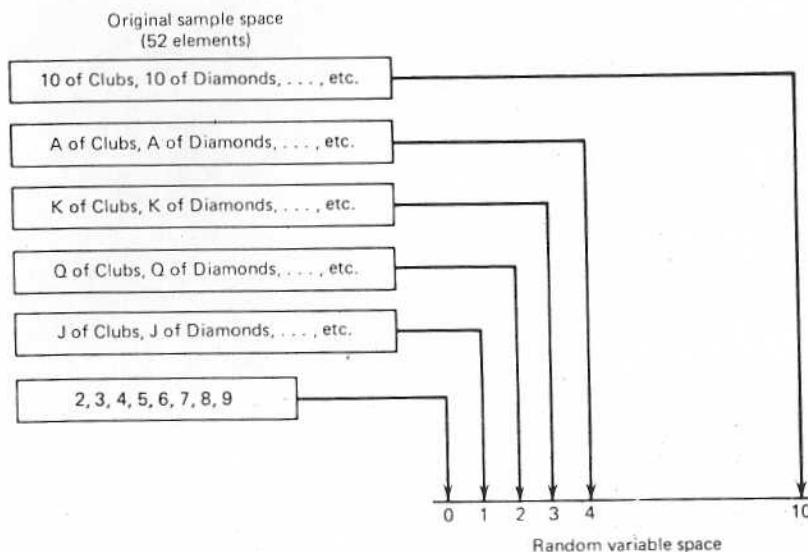


Figure 1.4. Mapping for pitch example.

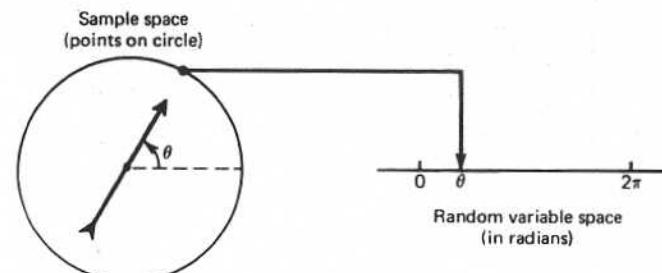


Figure 1.5. Mapping for spin-the-pointer example.

numbers in accordance with the above table. This is also shown in Fig. 1.4. Note that multiple points in the sample space map into the *same* number on the real line. This is perfectly legitimate. The mapping must not be ambiguous in going from the sample space to the real line; however, it may be ambiguous going the other way. That is, the mapping need not be one-to-one. ■

Example 1.14 In many games, the player spins a pointer that is mounted on a circular card of some sort and is free to spin about its center. This is depicted in Fig. 1.5 and the circular card is intentionally shown without any markings along its edge. Suppose we define the outcome of an experiment as the location on the periphery of the card at which the pointer stops. The sample space then consists of an infinite number of points along a circle. For analysis purposes, we might wish to identify each point in the sample space in terms of an angular coordinate measured in radians. The functional mapping that maps all points on a circle to corresponding points on the real line between 0 and 2π would then define an appropriate random variable. ■

1.7 PROBABILITY DISTRIBUTION AND DENSITY FUNCTIONS

When the sample space consists of a finite number of elements, the probability assignment can be made directly on the sample-space elements in accordance with what we feel to be reasonable in terms of likelihood of occurrence. This then defines probabilities for all events defined on the sample space. These probabilities, in turn, transfer directly to equivalent events in the random-variable space. The allowable realizations (i.e., real numbers) in the random-variable space are elementary equivalent events themselves, so the result is a probability associated with each allowable realization in the random-variable space. The sum of these probabilities must be unity, just as in the original sample space, but the distribution need not be the same. A continuation of Example 1.13, Section 1.6, will illustrate this.

Example 1.15 The mapping from the sample space to the real line for the pitch card game was shown in Fig. 1.4. Let us assign equal probabilities for all elements in the original sample space. The probabilities for the allowable realizations in the random-variable space would then be:

Random Variable Realization	Probability
0	$\frac{32}{52}$
1	$\frac{4}{52}$
2	$\frac{4}{52}$
3	$\frac{4}{52}$
4	$\frac{4}{52}$
10	$\frac{4}{52}$

Note that the probabilities are not distributed uniformly in the random-variable space. The end result of the mapping is a set of real numbers representing the possible realizations of the random variable and a corresponding set of probabilities that sum to unity. Once this correspondence has been established, the original sample space is usually ignored. ■

The random variable of Example 1.15 is an example of a *discrete* random variable in that its allowable realizations are discrete (i.e., countable) rather than continuous. The associated discrete set of probabilities is sometimes referred to as the *probability mass distribution* or simply *probability distribution*.

We also have occasion to work with continuous random variables. As a matter of fact, the usual electronic noise that is encountered in a wide variety of applications is of this type, that is, the voltage (or current) may assume a continuous range of values. The corresponding sample space then also contains an infinite number of points, so we cannot assign probabilities directly on the points of the sample space; this must be done on the defined events. We will continue the spin-the-pointer example of Section 1.6 (Example 1.14) to illustrate how this is done.

Let X denote a continuous random variable corresponding to the angular position of the pointer after it stops. Presumably, this could be any angle between 0 and 2π radians; therefore, the probability of *exactly* any particular position is zero. Thus, we assign a probability to the event that the pointer stops within a certain angular range, say between 0 and θ radians. If all positions are equally likely, it is reasonable to assign probabilities as follows:

$$P(X \leq \theta) = \begin{cases} 0, & \theta < 0 \\ \frac{1}{2\pi} \theta, & 0 \leq \theta \leq 2\pi \\ 1, & \theta > 2\pi \end{cases} \quad (1.7.1)$$

Note that the probability assignment is a function of the parameter θ and the function is sketched in Fig. 1.6. The linear portion of the function between 0 and 2π is due to the "equally likely" assumption.

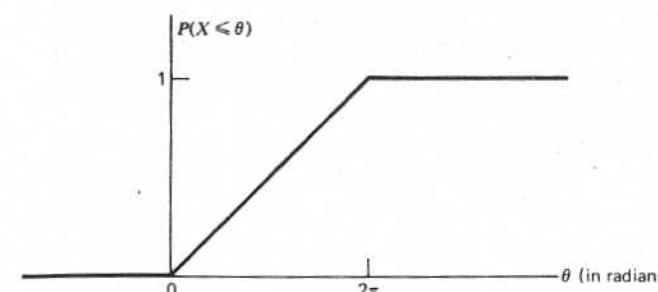


Figure 1.6. Probability distribution function for pointer example.

The function sketched in Fig. 1.6 is known as a *cumulative distribution function* (or just probability distribution function), and it simply describes the probability assignment as it reflects onto equivalent events in the random variable space. Specifically, the probability distribution function associated with the random variable X is defined as

$$F_X(\theta) = P(X \leq \theta) \quad (1.7.2)$$

where θ is a parameter representing a realization of X .

Just as in the discrete case, once this distribution is established in the random-variable space, the original sample space is usually ignored. It should be clear from the definition that a probability distribution function always has the following properties:

1. $F_X(\theta) \rightarrow 0$, as $\theta \rightarrow -\infty$.
2. $F_X(\theta) \rightarrow 1$, as $\theta \rightarrow \infty$.
3. $F_X(\theta)$ is a nondecreasing function of θ .

The information contained in the distribution function (e.g., Fig. 1.6) may also be presented in derivative form. Specifically, let $f_X(\theta)$ be defined as

$$f_X(\theta) = \frac{d}{d\theta} F_X(\theta) \quad (1.7.3)$$

The function $f_X(\theta)$ is known as the *probability density function* associated with the random variable X . The density function for the pointer example is shown in Fig. 1.7. From properties 1, 2, and 3 just cited for the distribution function, it should be apparent that the density function has the following properties:

1. $f_X(\theta)$ is a nonnegative function.
2. $\int_{-\infty}^{\infty} f_X(\theta) d\theta = 1$.

It should also be apparent from elementary calculus that the shaded area shown in Fig. 1.7 represents the probability that X lies between θ_1 and θ_2 . If θ_1 and θ_2 are separated by an infinitesimal amount, $\Delta\theta$, the area is approximately $f_X(\theta_1) \Delta\theta$, and thus we have the term *probability density*.

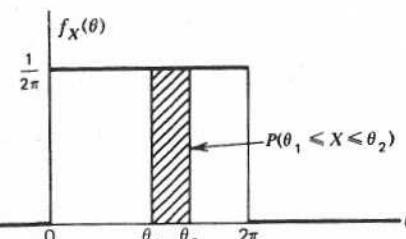


Figure 1.7. Probability density function for pointer example.

The probability density and distribution functions are alternative ways of describing the relative distribution of the random variable. Both functions are useful in random-variable analysis, and you should always keep in mind the derivative/integral relationship between the two. As a matter of notation, we will normally use an uppercase symbol for the distribution function and the corresponding lowercase symbol for the density function. The subscript in each case indicates the random variable being considered. The argument of the function is a dummy variable and may be most anything.

1.8 EXPECTATION, AVERAGES, AND CHARACTERISTIC FUNCTION

The idea of averaging is so commonplace that it may not seem worthy of elaboration. Yet, there are subtleties, especially as averaging relates to probability. Thus we need to formalize the notion of average.

Perhaps the first thing to note is that we always average over numbers and not "things." There is no such thing as the average of apples and oranges. When we compute a student's average grades, we do not average over *A*, *B*, *C*, etc.; instead, we average over numerical equivalents that have been arbitrarily assigned to each grade. Also, the quantities being averaged may or may not be governed by chance. In either case, random or deterministic, the average is just the sum of the numbers divided by the number of quantities being averaged. In the random case, the *sample average* or *sample mean* of a random variable *X* is defined as

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} \quad (1.8.1)$$

where X_1, X_2, \dots are sample realizations obtained from repeated trials of the chance situation under consideration. We will use the terms *average* and *mean* interchangeably, and the adjective *sample* serves as a reminder that we are averaging over a finite number of trials as in Eq. (1.8.1).

In the study of random variables we also like to consider the conceptual average that would occur for an infinite number of trials. This idea is basic to

the relative-frequency concept of probability. This hypothetical average is called *expected value* and is aptly named; it simply refers to what one would "expect" in the typical statistical situation. Beginning with discrete probability, imagine a random variable whose n possible realizations are x_1, x_2, \dots, x_n . The corresponding probabilities are p_1, p_2, \dots, p_n . If we have N trials, where N is large, we would expect approximately $p_1 N x_1$'s, $p_2 N x_2$'s, etc. Thus the sample average would be

$$\bar{X} \approx \frac{(p_1 N) x_1 + (p_2 N) x_2 + \dots + (p_n N) x_n}{N} \quad (1.8.2)$$

This suggests the following definition for expected value for the discrete probability case:

$$\text{Expected value of } X = E(X) = \sum_{i=1}^n p_i x_i \quad (1.8.3)$$

where n is the number of allowable values of the random variable X .

Similarly, for a continuous random variable X , we have

$$\text{Expected value of } X = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx \quad (1.8.4)$$

It should be mentioned that Eqs. (1.8.3) and (1.8.4) are definitions, and the arguments leading up to these definitions were presented to give a sensible rationale for the definitions, and not as a proof. We can use these same arguments for defining the expectation of a function of X , as well as for X . Thus we have

Discrete case:

$$E(g(X)) = \sum_i p_i g(x_i) \quad (1.8.5)$$

Continuous case:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad (1.8.6)$$

As an example of the use of Eq. (1.8.6), let the function $g(X)$ be X^k . Equation (1.8.6) [or its discrete counterpart Eq. (1.8.5)] then provides an expression for the *kth moment* of X , that is,

$$E(X^k) = \int_{-\infty}^{\infty} x^k f_X(x) dx \quad (1.8.7)$$

The second moment of X is of special interest, and it is given by

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx \quad (1.8.8)$$

The first moment is, of course, just the expectation of X , which is also known

as the *mean* or *average value* of X . Note when the term sample is omitted, we tacitly assume that we are referring to the hypothetical infinite-sample average.

We also have occasion to look at the second moment of X "about the mean." This quantity is called the *variance* of X and is defined as

$$\text{Variance of } X = E[(X - E(X))^2] \quad (1.8.9)$$

In a qualitative sense, the variance of X is a measure of the dispersion of X about its mean. Of course, if the mean is zero, the variance is identical to the second moment.

The expression for variance given by Eq. (1.8.9) can be reduced to a more convenient computational form by expanding the quantity within the brackets and then noting that the expectation of the sum is the sum of the expectations. This leads to

$$\begin{aligned} \text{Var } X &= E[X^2 - 2X \cdot E(X) + (E(X))^2] \\ &= E(X^2) - (E(X))^2 \end{aligned} \quad (1.8.10)$$

The square root of the variance is also of interest, and it has been given the name *standard deviation*, that is,

$$\text{Standard deviation of } X = \sqrt{\text{variance of } X} \quad (1.8.11)$$

Example 1.16 Let X be uniformly distributed in the interval $(0, 2\pi)$. This leads to the probability density function (see Example 1.14).

$$f_X(x) = \begin{cases} \frac{1}{2\pi}, & 0 \leq x < 2\pi \\ 0, & \text{elsewhere} \end{cases}$$

Find the mean, variance, and standard deviation of X .

The mean is just the expectation of X and is given by Eq. (1.8.4).

$$\begin{aligned} \text{Mean of } X = E(X) &= \int_0^{2\pi} x \cdot \frac{1}{2\pi} dx \\ &= \frac{1}{2\pi} \cdot \frac{x^2}{2} \Big|_0^{2\pi} = \pi \end{aligned} \quad (1.8.12)$$

Now that we have computed the mean, we are in a position to find the variance from Eq. (1.8.10).

$$\begin{aligned} \text{Var } X &= \int_0^{2\pi} x^2 \cdot \frac{1}{2\pi} dx - \pi^2 \\ &= \frac{4}{3} \pi^2 - \pi^2 \\ &= \frac{1}{3} \pi^2 \end{aligned} \quad (1.8.13)$$

The standard deviation is now just the square root of the variance:

$$\begin{aligned} \text{Standard deviation of } X &= \sqrt{\text{Var } X} \\ &= \sqrt{\frac{1}{3} \pi^2} = \frac{1}{\sqrt{3}} \pi \end{aligned} \quad (1.8.14)$$

The *characteristic function* associated with the random variable X is defined as

$$\psi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{j\omega x} dx \quad (1.8.15)$$

It can be seen that $\psi_X(\omega)$ is just the Fourier transform of the probability density function with a reversal of sign on ω . Thus the theorems (and tables) of Fourier transform theory can be used to advantage in evaluating characteristic functions and their inverses.

The characteristic function is especially useful in evaluating the moments of X . This can be demonstrated as follows. The moments of X may be written as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx \quad (1.8.16)$$

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &\vdots \\ &\text{etc.} \end{aligned} \quad (1.8.17)$$

Now consider the derivatives of $\psi_X(\omega)$ evaluated at $\omega = 0$.

$$\left[\frac{d\psi_X}{d\omega} \right]_{\omega=0} = \left[\int_{-\infty}^{\infty} jx f_X(x) e^{j\omega x} dx \right]_{\omega=0} = \int_{-\infty}^{\infty} jx f_X(x) dx \quad (1.8.18)$$

$$\begin{aligned} \left[\frac{d^2\psi_X}{d\omega^2} \right]_{\omega=0} &= \left[\int_{-\infty}^{\infty} (jx)^2 f_X(x) e^{j\omega x} dx \right]_{\omega=0} = \int_{-\infty}^{\infty} j^2 x^2 f_X(x) dx \\ &\vdots \\ &\text{etc.} \end{aligned} \quad (1.8.19)$$

It can be seen that

$$E(X) = \frac{1}{j} \left[\frac{d\psi_X}{d\omega} \right]_{\omega=0} \quad (1.8.20)$$

$$E(X^2) = \frac{1}{j^2} \left[\frac{d^2\psi_X}{d\omega^2} \right]_{\omega=0} \quad (1.8.21)$$

etc.

Thus, with the help of a table of Fourier transforms, you can often evaluate the moments without performing the integrations indicated in their definitions. [See Problems 1.24(b) and 2.28 for applications of the characteristic function.]

1.9 NORMAL OR GAUSSIAN RANDOM VARIABLES

The random variable X is called *normal* or *Gaussian* if its probability density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (x - m_X)^2 \right] \quad (1.9.1)$$

Note that this density function contains two parameters m_X and σ^2 . These are the random variable's mean and variance. That is, for the f_X specified by Eq. (1.9.1),

$$\int_{-\infty}^{\infty} x f_X(x) dx = m_X \quad (1.9.2)$$

and

$$\int_{-\infty}^{\infty} (x - m_X)^2 f_X(x) dx = \sigma^2 \quad (1.9.3)$$

Note that the normal density function is completely specified by assigning numerical values to the mean and variance. Thus a shorthand notation has come into common usage to designate a normal random variable. When we write

$$X \sim N(m_X, \sigma^2) \quad (1.9.4)$$

we mean X is normal with its mean given by the first argument in parentheses and its variance by the second argument. Also, as a matter of terminology, the terms normal and Gaussian are used interchangeably in describing normal random variables, and we will make no distinction between the two.

The normal density and distribution functions are sketched in Fig. 1.8a and 1.8b. Note that the density function is symmetric and peaks at its mean. Qualitatively, then, the mean is seen to be the most likely value, with values on either side of the mean gradually becoming less and less likely as the distance

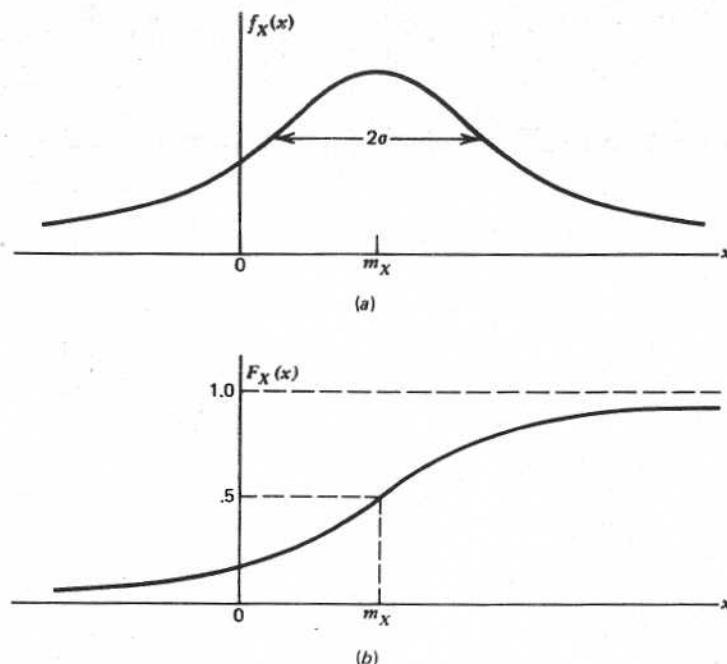


Figure 1.8. Normal density and distribution functions. (a) Normal density function. (b) Normal distribution function.

from the mean becomes larger. Since many natural random phenomena seem to exhibit this central-tendency property, at least approximately, the normal distribution is encountered frequently in applied probability. Recall that the variance is a measure of dispersion about the mean. Thus small σ corresponds to a sharp-peaked density curve, whereas large σ will yield a curve with a flat peak.

The normal distribution function is, of course, the integral of the density function:

$$F_X(x) = \int_{-\infty}^x f_X(u) du = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (u - m_X)^2 \right] du \quad (1.9.5)$$

Unfortunately, this integral cannot be represented in closed form in terms of elementary functions. Thus, its value must be tabulated as a function of the upper limit. A brief tabulation for zero mean and unity variance is given in Table 1.9. A quick glance at the table will show that the distribution function is very close to unity for values of the argument greater than 4.0. In our table, which was taken from Feller (4), the tabulation quits at about this point, and this is adequate for most purposes. However, in some instances, the difference between F_X and unity, be it ever so small, is very much of interest and we need to be able to compute it. An approximate formula for doing so for zero mean

and unity variance is given by

$$1 - F_X(x) \approx \frac{1}{\sqrt{2\pi}x} e^{-\frac{x^2}{2}} \quad (1.9.6)$$

Feller (4) may be consulted for a justification of this formula. The accuracy of Eq. (1.9.6) is only within a few percent for x equal to 4. It improves rapidly, though, as x increases.

Table 1.9 The Normal Density and Distribution Functions for Zero Mean and Unity Variance

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

x	$f_X(x)$	$F_X(x)$
.0	.398 942	.500 000
.1	.396 952	.539 828
.2	.391 043	.579 260
.3	.381 388	.617 911
.4	.368 270	.655 422
.5	.352 065	.691 462
.6	.333 225	.725 747
.7	.312 254	.758 036
.8	.289 692	.788 145
.9	.266 085	.815 940
1.0	.241 971	.841 345
1.1	.217 852	.864 334
1.2	.194 186	.884 930
1.3	.171 369	.903 200
1.4	.149 727	.919 243
1.5	.129 518	.933 193
1.6	.110 921	.945 201
1.7	.094 049	.955 435
1.8	.078 950	.964 070
1.9	.065 616	.971 283
2.0	.053 991	.977 250
2.1	.043 984	.982 136
2.2	.035 475	.986 097
2.3	.028 327	.989 276
2.4	.022 395	.991 802
2.5	.017 528	.993 790
2.6	.013 583	.995 339
2.7	.010 421	.996 533
2.8	.007 915	.997 445
2.9	.005 953	.998 134

Table 1.9 (continued)

x	$f_X(x)$	$F_X(x)$
3.0	.004 432	.998 650
3.1	.003 267	.999 032
3.2	.002 384	.999 313
3.3	.001 723	.999 517
3.4	.001 232	.999 663
3.5	.000 873	.999 767
3.6	.000 612	.999 841
3.7	.000 425	.999 892
3.8	.000 292	.999 928
3.9	.000 199	.999 952
4.0	.000 134	.999 968
4.1	.000 089	.999 979
4.2	.000 059	.999 987
4.3	.000 039	.999 991
4.4	.000 025	.999 995
4.5	.000 016	.999 997

It is worth mentioning that tables of normal distribution are presented in different ways in various references. For example, some tables present the one-sided area under the normal density curve from 0 to x , rather than from $-\infty$ to x as we have done in Table 1.9. Other tables do something similar by tabulating a function known as the error function (5), which is normalized differently than the usual distribution function. Thus a word of warning is in order. Be wary in the use of unfamiliar tables! An example illustrating the use of our brief tabulation in Table 1.9 is now in order.

Example 1.17 Let us say that the random variable X is normal with a mean of 1 and a variance of 4. In our shortened notation we could then write $X \sim N(1,4)$. Suppose we wish to find:

- (a) The value of the density function at its peak.
- (b) The probability that $X \geq 2$.
- (c) The probability that $0 \leq X \leq 2$.

We must first remember that the tabulated quantities are normalized for unity variance and zero mean. Thus the $f_X(x)$ of the tables is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{8}}$$

In part (a) we are asked to find the peak value of $\frac{1}{\sqrt{2\pi} \cdot 2} \exp \left[-\frac{1}{2 \cdot 4} (x-1)^2 \right]$. Clearly, this occurs where $x = 1$ and is $1/2 \sqrt{2\pi} \approx .199$. This could be obtained from Table 1.9 by noting the value of $f_X(x)$ for $x = 0$ and

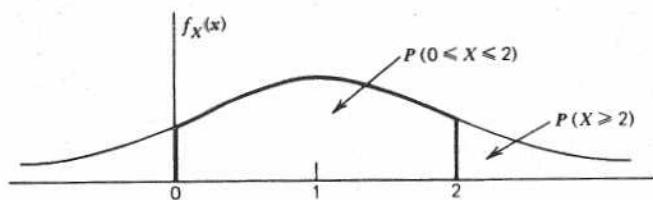


Figure 1.9. Key areas for Example 1.17.

dividing by 2. The division is necessary because σ (2 in this case) appears in the denominator of the normal density function. Of course, with modern pocket calculators, one does not need tables to evaluate the density function. Not so with the distribution function though. Most calculators are not programmed with an algorithm for this function, so tables are still necessary.

The areas involved in the solution of parts (b) and (c) are shown in Fig. 1.9. Formally, we wish to evaluate the following two integrals:

$$P(0 \leq X \leq 2) = \int_0^2 \frac{1}{\sqrt{2\pi} \cdot 2} \exp \left[-\frac{1}{2 \cdot 4} (u - 1)^2 \right] du \quad (1.9.7)$$

and

$$\begin{aligned} P(X \geq 2) &= \int_2^\infty \frac{1}{\sqrt{2\pi} \cdot 2} \exp \left[-\frac{1}{2 \cdot 4} (u - 1)^2 \right] du \\ &= 1 - \int_{-\infty}^2 \frac{1}{\sqrt{2\pi} \cdot 2} \exp \left[-\frac{1}{2 \cdot 4} (u - 1)^2 \right] du \end{aligned} \quad (1.9.8)$$

First we make the obvious change of variable:

$$v = \frac{u - 1}{2}$$

Then $du = 2 dv$, and the integrals become

$$P(0 \leq X \leq 2) = \int_{-1}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}v^2} dv \quad (1.9.9)$$

and

$$P(X \geq 2) = 1 - \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}v^2} dv \quad (1.9.10)$$

The tables can be used directly for the second term of Eq. (1.9.10), and therefore the answer for part (b) is

$$P(X \geq 2) = 1 - .691462 = .308538 \quad (1.9.11)$$

By taking advantage of symmetry, Eq. (1.9.9) can be written in the form

$$\begin{aligned} P(0 \leq X \leq 2) &= 2 \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}v^2} dv \\ &= 2 \left[\int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}v^2} dv - .5 \right] \end{aligned} \quad (1.9.12)$$

The tables may now be used to evaluate Eq. (1.10.12), and this gives the solution to part (c):

$$P(0 \leq X \leq 2) = 2[.691462 - .5] = .382924 \quad (1.9.13)$$

We can, of course, be less formal and take a more geometric approach in the solution of parts (b) and (c). This involves translating and rescaling the horizontal axis in accordance with the mean and variance of the random variable under consideration. The areas under the curve relative to the new axis then have a direct relationship to the quantities tabulated in Table 1.9. This is relatively obvious and will not be pursued further. ■

1.10 IMPULSIVE PROBABILITY DENSITY FUNCTIONS

In the case of the normal distribution, and many others, the probability associated with the random variable X is smoothly distributed over the real line from $-\infty$ to ∞ . The corresponding probability density function is then continuous, and the probability that any particular value of X , say x_0 , is realized is zero. This situation is common in physical problems, but we also have occasion to consider cases where the random variable has a mixture of discrete and smooth distribution. Rectification or any sort of hard limiting of noise leads to this situation, and an example will illustrate how this affects the probability density and distribution functions.

Example 1.18 Consider a simple half-wave rectifier driven by noise as shown in Fig. 1.10. For our purpose here, it will suffice to assume that the amplitude of the noise is normally distributed with zero mean. That is, if we were to sample the input at any particular time t_1 , the resultant sample is a

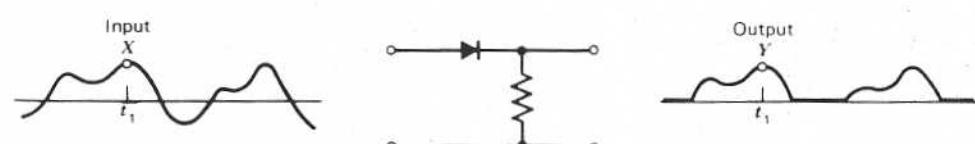


Figure 1.10. Half-wave rectifier driven by noise.

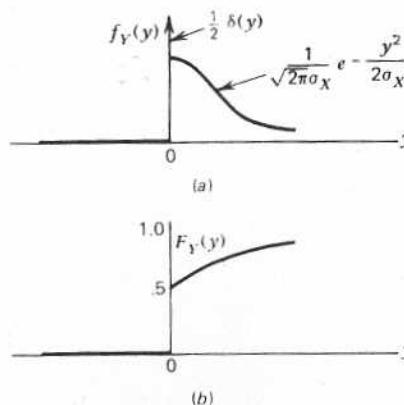


Figure 1.11. Output density and distribution functions for Example 1.18. (a) Probability density function for Y . (b) Probability distribution function for Y .

random variable, say X , whose distribution is $N(0, \sigma_X^2)$. The corresponding output sample is, of course, a different random variable; it will be denoted as Y .

Because of our assumption of normality, X will have probability density and distribution functions as shown in Fig. 1.8, but with $m_X = 0$. The sample space in this case may be thought of as all points on the real line, and the function defining the random variable X is just a one-to-one mapping of the sample space into X . Not so with Y though; all positive points in the sample space map one-to-one, but the negative points all map into zero because of the diode! This means that a total probability of $\frac{1}{2}$ in the sample space gets squeezed into a single point, zero, in the Y space. The effect of this on the density and distribution functions for Y is shown in Fig. 1.11a and 1.11b. Note that in order to have the area under the density function be $\frac{1}{2}$ at $y = 0$, we must have a Dirac delta (impulse) function at the origin. This, in turn, gives rise to a jump or discontinuity in the corresponding distribution function. It should be mentioned that at the point of discontinuity, the value of the distribution function is $\frac{1}{2}$. That is, the distribution function is continuous from the right and not from the left. This is due to the “equal to or less than . . .” statement in the definition of the probability distribution function (see Section 1.8). ■

1.11 MULTIPLE RANDOM VARIABLES

In subsequent chapters, we will frequently have occasion to deal with multiple random variables and their mutual relationships. Multiple random variables are often referred to as *multivariate* random variables, with *bivariate* being the special case of two variables. The various probabilistic relationships will be illustrated here for the bivariate case. The extension to three or more random variables is straightforward, so it will not be discussed specifically.

Let us first consider two *discrete* random variables X and Y . By discrete we mean that X and Y may take on discrete values x_i and y_j , where i and j are certain allowed integers. We will define the *joint probability distribution* (or *joint probability mass distribution*) as

$$p_{XY}(x_i, y_j) = P(X = x_i \text{ and } Y = y_j) \quad (1.11.1)$$

Note that this is not a cumulative distribution and, as an extra reminder, we will denote the discrete distribution with p rather than F . Just as in the case of joint events A and B (Section 1.4), the joint distribution of X and Y can be thought of as a two-dimensional array of probabilities, with each element in the array representing the probability of occurrence of a particular combination of X and Y . The sum of the numbers in the array obviously must be unity. Also, summing horizontally or vertically yields the marginal probabilities, just as in the “events” case. An example will illustrate these concepts.

Example 1.19 A sack contains 2 pennies, 1 nickel, and a dime. A coin is withdrawn at random and then replaced; then a second coin is withdrawn. The face value of the first draw (i.e., possible values are 1, 5, or 10) will be called random variable X , and the value of the second coin will be Y . We will assume that the outcome of the first draw does not affect the second in any way. The sample space in this case consists of 16 elements, and it, along with mapping into the bivariate random variable (X, Y) , is shown in Fig. 1.12. The two pennies are distinguished as Pen 1 and Pen 2. Note that the 16 elements of the sample space map into 9 2-tuples in the random-variable space. We have shown the 2-tuples in Fig. 1.12 as two numbers separated by commas. We could, of course,

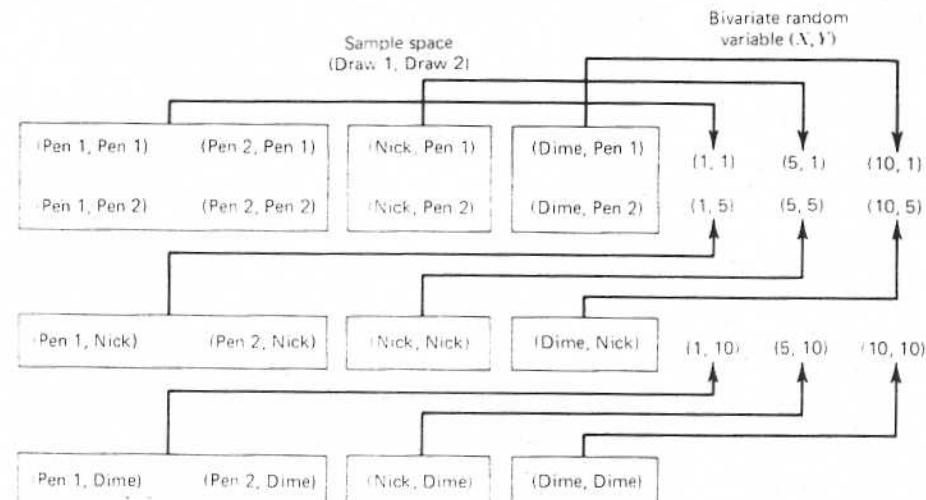


Figure 1.12. Sample space and mapping for Example 1.19.

Table 1.10 Joint and Marginal Probabilities for X and Y (X is First Draw and Y is Second Draw)

$X \backslash Y$	1	5	10	Marginal Probability $p(x_i)$	
1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$	
5	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$	
10	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$	
Marginal probability $p(y_j)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	Sum is unity when summed either horizontally or vertically	

have shown them as two-element column vectors; it's purely a matter of notation. If we assume that all elementary events in the sample space are equally likely, each should be assigned a probability of $\frac{1}{16}$. The joint probability distribution for X and Y is then as shown in Table 1.10. ■

Just as in Section 1.4, we can write certain relationships among the joint, marginal, and conditional probabilities for random variables X and Y .

Marginal (unconditional) probability:

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j) \quad (1.11.2)$$

$$p_Y(y_j) = \sum_i p_{XY}(x_i, y_j) \quad (1.11.3)$$

Conditional probability:

$$p_{X|Y} = \frac{p_{XY}}{p_Y} \quad (1.11.4)$$

$$p_{Y|X} = \frac{p_{XY}}{p_X} \quad (1.11.5)$$

Bayes rule:

$$p_{X|Y} = \frac{p_{Y|X}p_X}{p_Y} \quad (1.11.6)$$

The arguments in Eqs. (1.11.4) to (1.11.6) were omitted because the permissible x_i and y_j are obvious.

The discrete random variables X and Y are defined to be *statistically independent* if

$$p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j) \quad (1.11.7)$$

for all allowable x_i and y_j . As an example, the random variables in Example 1.19 are statistically independent because all nine probabilities in Table 1.10 satisfy the product relationship given by Eq. (1.11.7).

Let us now turn our attention to continuous random variables. Just as in the single-variable case, the relative distribution in the multivariate case must be described in terms of either a cumulative distribution function or a density function. Let X and Y be continuous random variables. The *joint cumulative distribution function* is defined* as

$$F_{XY}(x, y) = P(X \leq x \text{ and } Y \leq y) \quad (1.11.8)$$

Clearly, F_{XY} has the following properties:

$$(a) F_{XY}(-\infty, -\infty) = 0 \quad (1.11.9)$$

$$(b) F_{XY}(\infty, \infty) = 1 \quad (1.11.10)$$

$$(c) F_{XY} \text{ is nondecreasing in } x \text{ and } y \quad (1.11.11)$$

The joint density function of continuous random variables X and Y is given by

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad (1.11.12)$$

Note that there is an integral/derivative relationship between the cumulative distribution and density functions, just as in the single-variate case. Thus, to find the probability that a joint realization of X and Y will lie within a certain region R in the xy plane, we use the double integral formula

$$P(X \text{ and } Y \text{ lie within } R) = \iint_R f_{XY}(x, y) dx dy \quad (1.11.13)$$

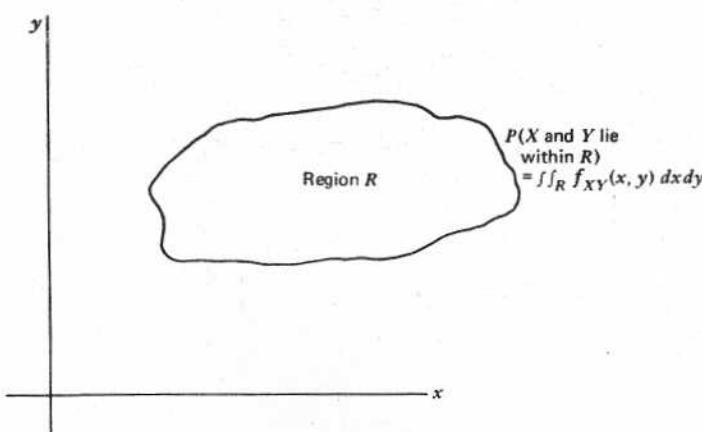
This is shown geometrically in Fig. 1.13. Of course, if the region R is a differential rectangle, as shown in Fig. 1.14, the probability of X and Y lying within the rectangle is

$$P(x_0 \leq X \leq x_0 + dx \text{ and } y_0 \leq Y \leq y_0 + dy) = f_{XY}(x_0, y_0)dx dy \quad (1.11.14)$$

Thus it should be apparent that f_{XY} has the usual meaning of density in a two-dimensional sense.

The *marginal* or *unconditional densities* are obtained in a manner similar to the discrete case, except that summation is replaced with integration. Thus

* The term "distribution" is a bit overworked in probability theory. The cumulative distribution of continuous random variables is much different than the probability distribution associated with discrete random variables. Thus we will append the word "cumulative" here to avoid confusion.

Figure 1.13. Region R in xy plane.

we have

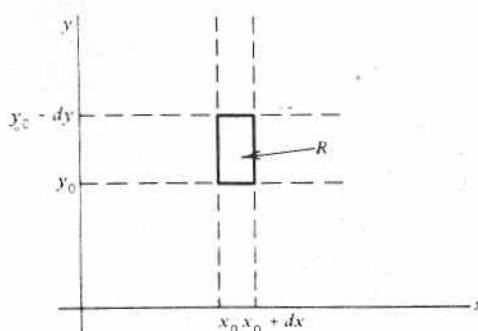
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy \quad (1.11.15)$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx \quad (1.11.16)$$

Equations (1.11.4) and (1.11.5) for discrete conditional probabilities can be applied to density functions for differential regions. This results in similar equations for conditional densities. The differential regions shown in Fig. 1.14 will be used in the following development. From the basic definition of conditional probability, we have

$$P(X \text{ is in strip } dx | Y \text{ is in strip } dy) = \frac{f_{XY}(x_0, y_0)dx dy}{f_Y(y_0) dy} \quad (1.11.17)$$

Figure 1.14. Differential area R in xy plane.

Cancelling the dy 's and noting the statement "Y is in strip dy " is approximately the same as saying "Y is equal to y_0 " leads to

$$P(x_0 \leq X \leq x_0 + dx | Y = y_0) = \left[\frac{f_{XY}(x_0, y_0)}{f_Y(y_0)} \right] dx \quad (1.11.18)$$

Note that the quantity in brackets in Eq. (1.11.18) has all the earmarks of a density function, and its probabilistic interpretation is on the left side of the equation. Thus we define conditional density as

$$f_{X|Y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)} \quad (1.11.19)$$

The functional dependence on the second argument y has been intentionally omitted in $f_{X|Y}$. This is to emphasize that $f_{X|Y}$ is a density function on x , not y [see Eq. (1.11.18)]. Of course, y does appear in $f_{X|Y}$ as a parameter, which may be thought of as a given, deterministic quantity. It is not the primary argument of $f_{X|Y}$ though.

The analogous definition of $f_{Y|X}$ is

$$f_{Y|X}(y) = \frac{f_{XY}(x, y)}{f_X(x)} \quad (1.11.20)$$

Once Eqs. (1.11.19) and (1.11.20) have been established, Bayes rule follows directly.

Bayes rule:

$$f_{X|Y}(x) = \frac{f_{Y|X}(y)f_X(x)}{f_Y(y)} \quad (1.11.21)$$

Statistical independence for continuous random variables X and Y is defined in the same manner as for discrete variables; that is, X and Y are statistically independent if

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad (1.11.22)$$

An example is now in order.

Example 1.20 Consider a dart-throwing game in which the target is a conventional xy coordinate system. The player aims each throw at the origin according to his or her best ability. Since there are many vagaries affecting each throw, we can expect a scatter in the hit pattern. Also, after some practice, the scatter should be unbiased, left-to-right and up-and-down. Let the coordinate of a hit be a bivariate random variable (X, Y) . In this example we would not expect the x coordinate to affect y in any way; therefore statistical independence of X and Y is a reasonable assumption. Also, because of the central tendency in X and Y , the assumption of normal distribution would appear to be

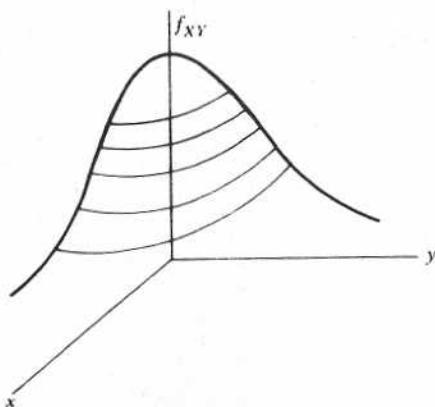


Figure 1.15. Sketch of one quadrant of a bivariate normal density function.

reasonable. Thus we assume the following unconditional densities:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \quad (1.11.23)$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} \quad (1.11.24)$$

The joint density is then given by

$$f_{XY}(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (1.11.25)$$

Equation (1.11.25) is the special case of a bivariate normal density function in which X and Y are independent, unbiased, and have equal variances. This is an important density function and it is sketched in Fig. 1.15. It is often de-

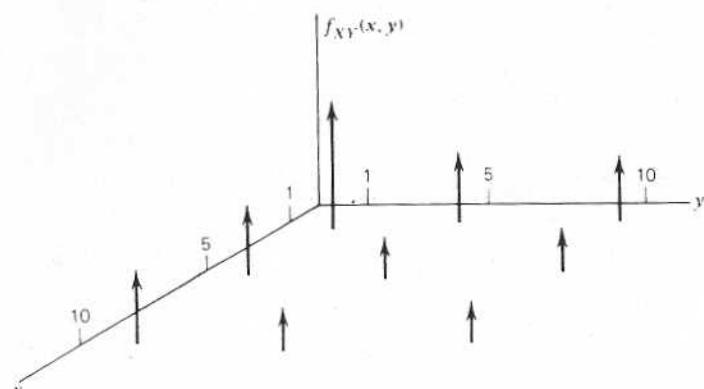


Figure 1.16. Impulsive joint probability density function for coin-drawing example. Example 1.19.

scribed as a smooth "hill-shaped" function and, for the special case here, the hill is symmetric in every respect. Thus the contours shown in Fig. 1.15 are exact circles. ■

The joint density function for a discrete bivariate random variable (X, Y) can be represented with impulsive-type functions, just as in the single-variate case. However, in the bivariate case, remember that volume (not area) represents probability. Thus volume-type impulse functions are appropriate. When X and Y are independent, the appropriate two-dimensional impulse function is obtained as a simple product of Dirac delta functions. For example, a unit volume impulse at the origin is given by $\delta(x)\delta(y)$. However, the situation is more complicated when X and Y are not independent, and we simply have to imagine appropriate volume-type impulsive functions. We will consider examples of this in subsequent chapters. The impulsive density function representation for the coin-drawing example, Example 1.19, is shown in Fig. 1.16.

1.12 CORRELATION, COVARIANCE, AND ORTHOGONALITY

The expectation of the product of two random variables X and Y is of special interest. In general, it is given by

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy \quad (1.12.1)$$

There is a special simplification of Eq. (1.12.1) that occurs when X and Y are independent. In this case, f_{XY} may be factored (see Eq. 1.11.22). Equation (1.12.1) then reduces to

$$E(XY) = \int_{-\infty}^{\infty} xf_X(x) dx \int_{-\infty}^{\infty} yf_Y(y) dy = E(X)E(Y) \quad (1.12.2)$$

If X and Y possess the property of Eq. (1.12.2), that is, the expectation of the product is the product of the individual expectations, they are said to be *uncorrelated*. Obviously, if X and Y are independent, they are also uncorrelated. However, the converse is not true, except in a few special cases (see Section 1.16).

As a matter of terminology, if

$$E(XY) = 0 \quad (1.12.3)$$

X and Y are said to be *orthogonal*.

The *covariance* of X and Y is also of special interest, and it is defined as

$$\text{Cov of } X \text{ and } Y = E[(X - m_X)(Y - m_Y)] \quad (1.12.4)$$

With the definition of Eq. (1.12.4) we can now define the *correlation coefficient* for two random variables as

$$\begin{aligned}\text{Correlation coefficient } \rho &= \frac{\text{Cov of } X \text{ and } Y}{\sqrt{\text{Var } X} \sqrt{\text{Var } Y}} \\ &= \frac{E[(X - m_X)(Y - m_Y)]}{\sqrt{\text{Var } X} \sqrt{\text{Var } Y}}\end{aligned}\quad (1.12.5)$$

The correlation coefficient is a normalized measure of the degree of correlation between two random variables, and the normalization is such that ρ always lies within the range $-1 \leq \rho \leq 1$. This will be demonstrated (not proved) by looking at three special cases:

1. $Y = X$ (maximum positive correlation):

When $Y = X$, Eq. (1.12.5) becomes

$$\rho = \frac{E[(X - m_X)(X - m_X)]}{\sqrt{E(X - m_X)^2} \sqrt{E(X - m_X)^2}} = 1$$

2. $Y = -X$ (maximum negative correlation):

When $Y = -X$, Eq. (1.12.5) becomes

$$\rho = \frac{E[(X - m_X)(-X + m_X)]}{\sqrt{E(X - m_X)^2} \sqrt{E(-X + m_X)^2}} = -1$$

3. X and Y uncorrelated, that is, $E(XY) = E(X)E(Y)$:

Expanding the numerator of Eq. (1.12.5) yields

$$\begin{aligned}E(XY - m_X Y - m_Y X + m_X m_Y) \\ = E(XY) - m_X E(Y) - m_Y E(X) + m_X m_Y \\ = m_X m_Y - m_X m_Y - m_Y m_X + m_X m_Y = 0\end{aligned}$$

Thus $\rho = 0$.

We have now examined the extremes of positive and negative correlation and zero correlation; there can be all shades of gray in between. [For further details, see Papoulis (6)].

1.13 SUM OF INDEPENDENT RANDOM VARIABLES AND TENDENCY TOWARD NORMAL DISTRIBUTION

Since we frequently need to consider additive combinations of independent random variables, this will now be examined in some detail. Let X and Y be independent random variables with probability density functions $f_X(x)$ and

$f_Y(y)$. Define another random variable Z as the sum of X and Y :

$$Z = X + Y \quad (1.13.1)$$

Given the density functions of X and Y , we wish to find the corresponding density of Z .

Let z be a particular realization of the random variable Z , and think of z as being fixed. Now consider all possible realizations of X and Y that yield z . Clearly, they satisfy the equation

$$x + y = z \quad (1.13.2)$$

and the locus of points in the x, y plane is just a straight line, as shown in Fig. 1.17.

Next, consider an incremental perturbation of z to $z + dz$, and again consider the locus of realizations of X and Y that will yield $z + dz$. This locus is also a straight line, and it is shown as the upper line in Fig. 1.17. It should be apparent that all x and y lying within the differential strip between the two lines map into points between z and $z + dz$ in the z space. Therefore,

$$\begin{aligned}P(z \leq Z \leq z + dz) &= P(x \text{ and } y \text{ lie in differential strip}) \\ &= \iint_{\text{Diff. strip}} f_X(x)f_Y(y) dx dy\end{aligned}\quad (1.13.3)$$

But, within the differential strip, y is constrained to x according to

$$y = z - x \quad (1.13.4)$$

Also, since the strip is only of differential width, the double integral of Eq. (1.13.3) reduces to a single integral. Choosing x as the variable of integration and noting that $dy = dz$ leads to

$$P(z \leq Z \leq z + dz) = \left[\int_{-\infty}^z f_X(x)f_Y(z - x) dx \right] dz \quad (1.13.5)$$

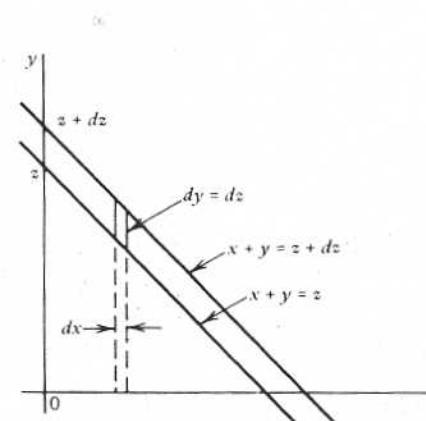


Figure 1.17. Differential strip used for deriving $f_Z(z)$.

It is now apparent from Eq. (1.13.5) that the quantity within the brackets is the desired probability density function for Z . Thus

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx \quad (1.13.6)$$

It is of interest to note that the integral on the right side of Eq. (1.13.6) is a convolution integral. Thus, from Fourier transform theory, we can then write

$$\mathcal{F}[f_Z] = \mathcal{F}[f_X] \cdot \mathcal{F}[f_Y] \quad (1.13.7)$$

where $\mathcal{F}[\cdot]$ denotes "Fourier transform of $[\cdot]$." We now have two ways of evaluating the density of Z : (1) We can evaluate the convolution integral directly, or (2) we can first transform f_X and f_Y , then form the product of the transforms, and finally invert the product to get f_Z . Examples that illustrate each of these methods follow.

Example 1.21 Let X and Y be independent random variables with identical rectangular density functions as shown in Fig. 1.18a. We wish to find the density function for their sum, which we will call Z .

Note first that the density shown in Fig. 1.18a has even symmetry. Thus $f_Y(z-x) = f_Y(x-z)$. The convolution integral expression of Eq. (1.13.6) is then the integral of a rectangular pulse multiplied by a similar pulse shifted to the right amount z . When $z > 2$ or $z < -2$ there is no overlap in the pulses so their product is zero. When $-2 \leq z \leq 0$ there is a nontrivial overlap which increases linearly beginning at $z = -2$ and reaching a maximum at $z = 0$. The convolution integral then increases accordingly as shown in Fig. 1.18b. A similar

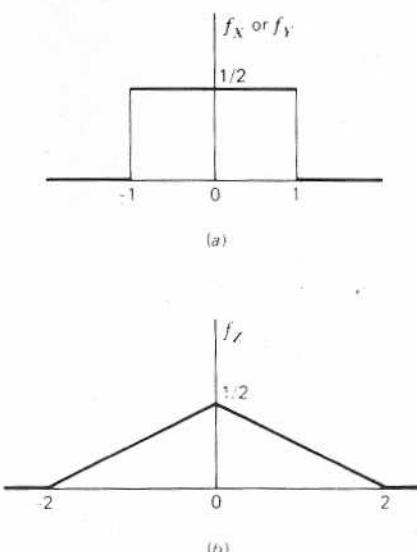


Figure 1.18. Probability density functions for X and Y and their sum. (a) Density function for both X and Y . (b) Density function for Z , where $Z = X + Y$.

argument may be used to show that $f_Z(z)$ decreases linearly in the interval where $0 \leq z \leq 2$. This leads to the triangular density function of Fig. 1.18b.

We can go one step further now and look at the density corresponding to the sum of three random variables. Let W be defined as

$$W = X + Y + V \quad (1.13.8)$$

where X , Y , and V are mutually independent and have identical rectangular densities as shown in Fig. 1.18a. We have already worked out the density for $X + Y$, so the density of W is the convolution of the two functions shown in Fig. 1.18a and 1.18b. We will leave the details of this as an exercise, and the result is shown in Fig. 1.19. Each of the segments labeled 1, 2, and 3 is an arc of a parabola. Notice the smooth central tendency. With a little imagination one can see a similarity between this and a zero-mean normal density curve. If we were to go another step and convolve a rectangular density with that of Fig. 1.19, we would get the density for the sum of four independent random variables. The resulting function would consist of connected segments of cubic functions extending from -4 to $+4$. Its appearance, though not shown, would resemble the normal curve even more than that of Fig. 1.19. And on and on—each additional convolution results in a curve that resembles the normal curve more closely than the preceding one.

This simple example is intended to demonstrate (not prove) that a superposition of independent random variables always tends toward normality, regardless of the distribution of the individual random variables contributing to the sum. This is known as the *central limit theorem* of statistics. It is a most remarkable theorem, and its validity is subject to only modest restrictions (3). In engineering applications the noise we must deal with is frequently due to a superposition of many small contributions. When this is so, we have good reason to make the assumption of normality. The central limit theorem says to do just that. Thus we have here one of the reasons for our seemingly exaggerated interest in normal random variables—they are a common occurrence in nature. ■

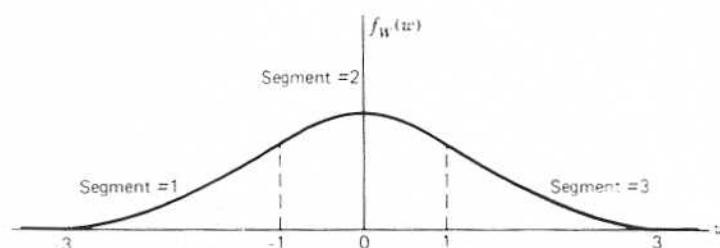


Figure 1.19. Probability density for the sum of three independent random variables with identical rectangular density functions.

Example 1.22 Let X and Y be independent normal random variables with zero means and variances σ_X^2 and σ_Y^2 . We wish to find the probability density function for the sum of X and Y , which will again be denoted as Z . For variety, we illustrate the Fourier transform approach. The explicit expressions for f_X and f_Y are

$$f_X(t) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-t^2/2\sigma_X^2} \quad (1.13.9)$$

$$f_Y(t) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-t^2/2\sigma_Y^2} \quad (1.13.10)$$

Note that we have used t as the dummy argument of the functions. It is of no consequence because it is integrated out in the transformation to the ω domain. Using Fourier transform tables, we find the transforms of f_X and f_Y to be

$$\mathcal{F}[f_X] = e^{-\sigma_X^2\omega^2/2} \quad (1.13.11)$$

$$\mathcal{F}[f_Y] = e^{-\sigma_Y^2\omega^2/2} \quad (1.13.12)$$

Forming their product yields

$$\mathcal{F}[f_X]\mathcal{F}[f_Y] = e^{-(\sigma_X^2+\sigma_Y^2)\omega^2/2} \quad (1.13.13)$$

Then the inverse gives the desired f_Z :

$$\begin{aligned} f_Z(z) &= \mathcal{F}^{-1}[e^{-(\sigma_X^2+\sigma_Y^2)\omega^2/2}] \\ &= \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}} e^{-z^2/2(\sigma_X^2+\sigma_Y^2)} \end{aligned} \quad (1.13.14)$$

Note that the density function for Z is also normal in form, and its variance is given by

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 \quad (1.13.15)$$

The summation of any number of random variables can always be thought of as a sequence of summing operations on two variables; therefore, it should be clear that summing any number of independent normal random variables leads to a normal random variable. This rather remarkable result can be generalized further to include the case of dependent normal random variables, which we will discuss later. ■

1.14 TRANSFORMATION OF RANDOM VARIABLES

A mathematical transformation that takes one set of variables (say, inputs) into another set (say, outputs) is a common situation in systems analysis. Let us begin with a simple single-input, single-output situation where the input-output

relationship is governed by the algebraic equation

$$y = g(x) \quad (1.14.1)$$

Here we are interested in random inputs, so think of x as a realization of the input random variable X , and y as the corresponding realization of the output Y . Assume we know the probability density function for X , and would like to find the corresponding density for Y . It is tempting to simply replace x in $f_X(x)$ with its equivalent in terms of y and pass it off at that. However, it is not quite that simple, as will be seen presently.

First, let us assume that the transformation $g(x)$ is one-to-one for all permissible x . By this, we mean that the functional relationship given by Eq. (1.14.1) can be reversed, and x can be written uniquely as a function of y . Let the "reverse" relationship be

$$x = h(y) \quad (1.14.2)$$

The probabilities that X and Y lie within corresponding differential regions must be equal. That is,

$$P(X \text{ is between } x \text{ and } x+dx) = P(Y \text{ is between } y \text{ and } y+dy) \quad (1.14.3)$$

or

$$\int_x^{x+dx} f_X(u) du = \begin{cases} \int_y^{y+dy} f_Y(u) du, & \text{for } dy \text{ positive} \\ - \int_y^{y+dy} f_Y(u) du, & \text{for } dy \text{ negative} \end{cases} \quad (1.14.4)$$

One of the subtleties of this problem should now be apparent from Eq. (1.14.4). If positive dx yields negative dy (i.e., a negative derivative), the integral of f_Y must be taken from $y + dy$ to y in order to yield a positive probability. This is the equivalent of interchanging the limits and reversing the sign as shown in Eq. (1.14.4).

The differential equivalent of Eq. (1.14.4) is

$$f_X(x) dx = f_Y(y)|dy| \quad (1.14.5)$$

where we have tacitly assumed dx to be positive. Also, x is constrained to be $h(y)$. Thus we have

$$f_Y(y) = \left| \frac{dx}{dy} \right| f_X(h(y)) \quad (1.14.6)$$

or, equivalently,

$$f_Y(y) = |h'(y)| f_X(h(y)) \quad (1.14.7)$$

where $h'(y)$ indicates the derivative of h with respect to y . Two examples will now be presented.

Example 1.23 Find the appropriate output density functions for the case where the input X is $N(0, \sigma_X^2)$ and the transformation is

$$(a) y = Kx \quad (K \text{ is a given constant}) \quad (1.14.8)$$

$$(b) y = x^3 \quad (1.14.9)$$

$$(c) y = x^2 \quad (1.14.10)$$

We begin with the scale-factor transformation indicated by Eq. (1.14.8). We first solve for x in terms of y and then form the derivative. Thus,

$$x = \frac{1}{K}y \quad (1.14.11)$$

$$\left| \frac{dx}{dy} \right| = \left| \frac{1}{K} \right| \quad (1.14.12)$$

We can now obtain the equation for f_Y from Eq. (1.14.6). The result is

$$f_Y(y) = \frac{1}{|K|} \cdot \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp \left[-\frac{(y/K)^2}{2\sigma_X^2} \right] \quad (1.14.13)$$

Or, rewriting Eq. (1.14.13) in standard normal form yields

$$f_Y(y) = \frac{1}{\sqrt{2\pi(K\sigma_X)^2}} \exp \left[-\frac{y^2}{2(K\sigma_X)^2} \right] \quad (1.14.14)$$

It can now be seen that transforming a zero-mean normal random variable with a simple scale factor yields another normal random variable with a corresponding scale change in its standard deviation. It is important to note that normality is preserved in a linear transformation.

Next, consider part (b). This transformation is also one-to-one, so solving for x yields

$$x = \sqrt[3]{y} \quad (1.14.15)$$

In Eq. (1.14.15) we take the positive real root for $y > 0$ and the negative real root for $y < 0$. The derivative of x is then

$$\frac{dx}{dy} = \frac{1}{3} y^{-2/3} \quad (1.14.16)$$

The quantity $y^{2/3}$ can be written as $(y^{1/3})^2$, so $y^{2/3}$ is always positive, provided $y^{1/3}$ is real. This is consistent with the geometric interpretation of $y = x^3$, which always has a nonnegative slope. The density function for Y is then

$$f_Y(y) = \frac{1}{3y^{2/3}} \cdot \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-(y^{1/3})^2/2\sigma_X^2} \quad (1.14.17)$$

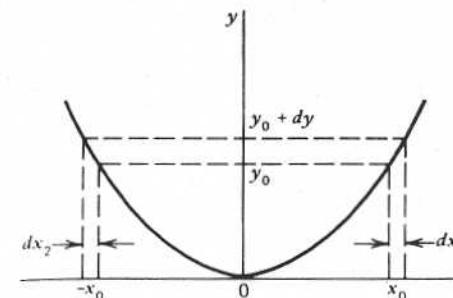


Figure 1.20. Sketch of $y = x^2$.

Since this does not reduce to normal form, we have here an example of a nonlinear transformation that converts a normal random variable to non-Gaussian form.

The transformation $y = x^2$ for part (c) is sketched in Fig. 1.20. It is obvious from the sketch that two values of x yield the same y . Thus the function is not one-to-one, which violates one of the assumptions in deriving Eq. (1.14.6). The problem is solvable, though; we simply must go back to fundamentals and derive a new $f_Y(y)$ relationship. Note that we can consider $y = x^2$ in terms of two branches. These will be defined as

$$x = \begin{cases} \sqrt{y}, & x \geq 0 \text{ (Branch 1)} \\ -\sqrt{y}, & x < 0 \text{ (Branch 2)} \end{cases} \quad (1.14.18)$$

Now think of perturbing y a positive differential amount dy as shown in Fig. 1.20. This results in two corresponding differential regions on the x axis. Thus we have the following relationship for probabilities:

$$P(y_0 \leq Y \leq y_0 + dy) = P(x_0 \leq X \leq x_0 + dx_1) + P(-x_0 + dx_2 \leq X \leq -x_0) \quad (1.14.19)$$

Note that dx_1 (Branch 1) is positive and dx_2 (Branch 2) is negative. Next, Eq. (1.14.19) can be rewritten in terms of integrals as

$$\int_{y_0}^{y_0 + dy} f_Y(u) du = \int_{x_0}^{x_0 + dx_1} f_X(v) dv + \int_{-x_0 + dx_2}^{-x_0} f_X(v) dv \quad (1.14.20)$$

We now note that

$$dx_1 = \frac{1}{2\sqrt{y_0}} dy \quad (1.14.21)$$

$$dx_2 = -\frac{1}{2\sqrt{y_0}} dy$$

Substituting Eq. (1.14.21) into Eq. (1.14.20) and letting the range of integration

be incremental yields

$$f_Y(y_0) dy = f_X(\sqrt{y_0}) \cdot \frac{1}{2\sqrt{y_0}} dy + f_X(\sqrt{y_0}) \cdot \frac{1}{2\sqrt{y_0}} dy, \quad y_0 \geq 0 \quad (1.14.22)$$

Now, since y_0 can be any y greater than zero, Eq. (1.14.22) reduces to

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}), \quad y \geq 0 \quad (1.14.23)$$

From the equation $y = x^2$, we see that no real values of x map into negative y . Therefore, $f_Y(y) = 0$ for negative y . The final f_Y is then

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{y}} f_X(\sqrt{y}), & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (1.14.24)$$

We can now apply Eq. (1.14.24) to the case where X is normal with zero mean. This yields

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_X \cdot \sqrt{y}} e^{-y/2\sigma_X^2}, & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (1.14.25)$$

Thus, we see another example of a nonlinear transformation leading to a non-normal random variable. ■

We also have occasion to deal with transformations of multiple random variables. Consider a bivariate transformation example. Since the extension to higher-order transformations is fairly obvious, only the bivariate case will be discussed in detail.

Example 1.24 In the target-shooting example (Example 1.20), the hit location was described in terms of rectangular coordinates. This led to the joint probability density function

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (1.14.26)$$

This is a special case of the bivariate normal density function where the two variates are independent and have zero means and equal variances. Suppose we wish to find the corresponding density in terms of polar coordinates r and θ . Formally, then, we wish to define new random variables R and Θ such that pairwise realizations (x, y) transform to (r, θ) in accordance with

$$\begin{aligned} r &= \sqrt{x^2 + y^2}, \quad r \geq 0 \\ \theta &= \tan^{-1} \frac{y}{x}, \quad 0 \leq \theta < 2\pi \end{aligned} \quad (1.14.27)$$

Or, equivalently,

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned} \quad (1.14.28)$$

We wish to find $f_{R\Theta}(r, \theta)$ and the unconditional density functions $f_R(r)$ and $f_\Theta(\theta)$.

The probability that a hit will lie within an area bounded by a closed contour C is given by

$$P(\text{Hit lies within } C) = \iint_{\substack{\text{area} \\ \text{enclosed} \\ \text{by } C}} f_{XY}(x, y) dx dy \quad (1.14.29)$$

We know from multivariable calculus (8) that the double integral in Eq. (1.14.29) can also be evaluated in the r, θ coordinate frame as

$$\iint_{\substack{\text{Region} \\ \text{enclosed} \\ \text{by } C}} f_{XY}(x, y) dx dy = \iint_{\substack{\text{Region} \\ \text{enclosed} \\ \text{by } C'}} f_{XY}(x(r, \theta), y(r, \theta)) |J\left(\frac{x, y}{r, \theta}\right)| dr d\theta \quad (1.14.30)$$

where C' is the contour in the r, θ plane corresponding to C in the x, y plane. That is, points within C map into points within C' . (Note that it is immaterial as to how we draw the "picture" in the r, θ coordinate frame. For example, if we choose to think of r and θ as just another set of Cartesian coordinates for plotting purposes, C' becomes a distortion of C .) The J quantity in Eq. (1.14.30) is the Jacobian of the transformation, and it is defined as

$$J\left(\frac{x, y}{r, \theta}\right) = \text{Det} \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} \end{bmatrix} \quad (1.14.31)$$

The vertical bars around J in Eq. (1.14.30) indicate absolute magnitude. We can argue now that if Eq. (1.14.30) is true for regions in general, it must also be true for differential regions. Let the differential region in the r, θ domain be bounded by r and $r + dr$ in one direction and by θ and $\theta + d\theta$ in the other. If it helps, think of plotting r and θ as Cartesian coordinates. The differential region in the r, θ domain is then rectangular, and the corresponding one in the x, y domain is a curvilinear differential rectangle (see Fig. 1.21). Now, by the very definition of joint density, the quantity multiplying $dr d\theta$ in Eq. (1.14.30) is seen to be the desired density function. That is,

$$f_{R\Theta}(r, \theta) = f_{XY}(x(r, \theta), y(r, \theta)) |J\left(\frac{x, y}{r, \theta}\right)| \quad (1.14.32)$$

In the transformation of this example,

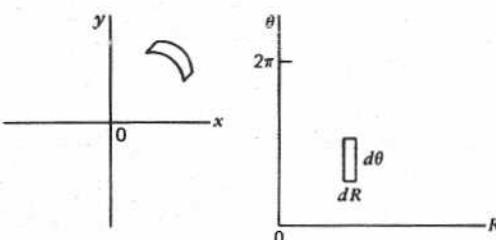


Figure 1.21. Corresponding differential regions for transformation of Example 1.24.

$$J\left(\frac{x}{r}, \theta\right) = \text{Det} \begin{bmatrix} \cos \theta & \sin \theta \\ -r \sin \theta & r \cos \theta \end{bmatrix} = r \quad (1.14.33)$$

Since the radial coordinate r is always positive, $|J| = r$. We can now substitute Eqs. (1.14.26) and (1.14.33) into (1.14.32) and obtain

$$\begin{aligned} f_{R\theta}(r, \theta) &= r \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(r \cos \theta)^2 + (r \sin \theta)^2}{2\sigma^2}\right] \quad (1.14.34) \\ &= \frac{r}{2\pi\sigma^2} e^{-r^2/2\sigma^2} \end{aligned}$$

Note that the density function of this example has no explicit dependence on θ . In other words, all angles between 0 and 2π are equally likely, which is what we would expect in the target-throwing experiment.

We get the unconditional density functions by integrating $f_{R\theta}$ with respect to the appropriate variables. That is,

$$\begin{aligned} f_R(r) &= \int_0^{2\pi} f_{R\theta}(r, \theta) d\theta = \frac{r}{2\pi\sigma^2} e^{-r^2/2\sigma^2} \int_0^{2\pi} d\theta \quad (1.14.35) \\ &= \frac{r}{\sigma^2} e^{-r^2/2\sigma^2} \end{aligned}$$

and

$$f_\theta(\theta) = \int_0^\infty f_{R\theta}(r, \theta) dr = \begin{cases} \frac{1}{2\pi}, & 0 \leq \theta < 2\pi \\ 0, & \text{otherwise} \end{cases} \quad (1.14.36)$$

The single-variate density function given by Eq. (1.14.35) is called the *Rayleigh density function*. It is of considerable importance in applied probability, and it is sketched in Fig. 1.22. It is easily verified that the mode (peak value) of the Rayleigh density is equal to standard deviation of the x and y normal random variables from which it was derived. Thus we see that similar indepen-

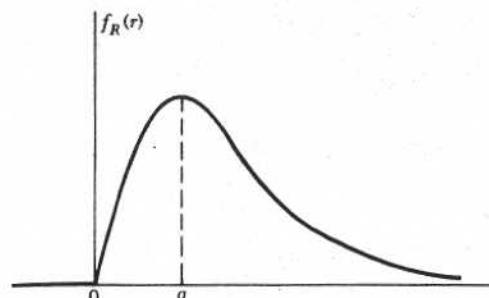


Figure 1.22. Rayleigh density function.

dent, zero-mean normal densities in the x, y domain correspond to Rayleigh and uniform densities in the r, θ domain. ■

1.15 MULTIVARIATE NORMAL DENSITY FUNCTION

In Sections 1.9 and 1.11 examples of the single and bivariate normal density functions were presented. We work so much with normal random variables that we need to elaborate further and develop a general form for the n -dimensional normal density function. One can write out the explicit equations for the single and bivariate cases without an undue amount of "clutter" in the equations. However, beyond this, matrix notation is a virtual necessity. Otherwise, the explicit expressions are completely unwieldy.

Consider a set of n random variables X_1, X_2, \dots, X_n (also called variates). We define a vector random variable \mathbf{X} as*

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad (1.15.1)$$

In general, the components of \mathbf{X} may be correlated and have nonzero means. We denote the respective means as m_1, m_2, \dots, m_n , and thus we define a

* Note that uppercase \mathbf{X} denotes a column vector in this section. This is a departure from the usual notation of matrix theory, but it is necessitated by a desire to be consistent with the previous uppercase notation for random variables. The reader will simply have to remember this minor deviation in matrix notation. It appears in this section and also occasionally in Chapter 2.

mean vector \mathbf{m} as

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} \quad (1.15.2)$$

Also, if x_1, x_2, \dots, x_n is a set of realizations of X_1, X_2, \dots, X_n , we can define a vector realization of \mathbf{X} as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (1.15.3)$$

We next define a matrix that describes the variances and correlation structure of the n variates. The *covariance matrix* for \mathbf{X} is defined as

$$\mathbf{C} = \begin{bmatrix} E[(X_1 - m_1)^2] & E[(X_1 - m_1)(X_2 - m_2)] & \dots \\ E[(X_2 - m_2)(X_1 - m_1)] & E[(X_2 - m_2)^2] & \dots \\ \vdots & \vdots & \ddots \\ E[(X_n - m_n)(X_{n-1} - m_{n-1})] & \dots & E[(X_n - m_n)^2] \end{bmatrix} \quad (1.15.4)$$

The terms along the major diagonal of \mathbf{C} are seen to be the variances of the variates, and the off-diagonal terms are the covariances.

The random variables X_1, X_2, \dots, X_n are said to be *jointly normal* or *jointly Gaussian* if their joint probability density function is given by

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}|^{1/2}} \exp \left\{ -\frac{1}{2} [(\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})] \right\} \quad (1.15.5)$$

where \mathbf{x} , \mathbf{m} , and \mathbf{C} are defined by Eqs. (1.15.2) to (1.15.4) and $|\mathbf{C}|$ is the determinant of \mathbf{C} . "Super -1" and "super T" denote matrix inverse and transpose, respectively. Note that the defining function for f_X is scalar and is a function of x_1, x_2, \dots, x_n when written out explicitly. We have shortened the indicated functional dependence to \mathbf{x} just for compactness in notation. Also note that \mathbf{C}^{-1} must exist in order for f_X to be properly defined by Eq. (1.15.5). Thus \mathbf{C} must be nonsingular. More will be said of this later.

Clearly, Eq. (1.15.5) reduces to the standard normal form for the single variate case. For the bivariate case, we may write out f_X explicitly in terms of x_1 and x_2 without undue difficulty. Proceeding to do this, we have

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} \quad (1.15.6)$$

and

$$\begin{aligned} \mathbf{C} &= \begin{bmatrix} E[(X_1 - m_1)^2] & E[(X_1 - m_1)(X_2 - m_2)] \\ E[(X_1 - m_1)(X_2 - m_2)] & E[(X_2 - m_2)^2] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \end{aligned} \quad (1.15.7)$$

The second form for \mathbf{C} in Eq. (1.15.7) follows directly from the definitions of variance and correlation coefficient. The determinant of \mathbf{C} and its inverse are given by

$$|\mathbf{C}| = \begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} = (1 - \rho^2)\sigma_1^2\sigma_2^2 \quad (1.15.8)$$

$$\mathbf{C}^{-1} = \begin{bmatrix} \frac{\sigma_2^2}{|\mathbf{C}|} & -\frac{\rho\sigma_1\sigma_2}{|\mathbf{C}|} \\ -\frac{\rho\sigma_1\sigma_2}{|\mathbf{C}|} & \frac{\sigma_1^2}{|\mathbf{C}|} \end{bmatrix} = \begin{bmatrix} \frac{1}{(1 - \rho^2)\sigma_1^2} & \frac{-\rho}{(1 - \rho^2)\sigma_1\sigma_2} \\ \frac{-\rho}{(1 - \rho^2)\sigma_1\sigma_2} & \frac{1}{(1 - \rho^2)\sigma_2^2} \end{bmatrix} \quad (1.15.9)$$

Substituting Eqs. (1.15.8) and (1.15.9) into Eq. (1.15.5) then yields the desired density function in terms of x_1 and x_2 .

$$f_{X_1 X_2}(x_1, x_2) =$$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1 - m_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - m_1)(x_2 - m_2)}{\sigma_1\sigma_2} + \frac{(x_2 - m_2)^2}{\sigma_2^2} \right] \right\} \quad (1.15.10)$$

It should be clear in Eq. (1.15.10) that $f_{X_1 X_2}(x_1, x_2)$ is intended to mean the same as $f_X(\mathbf{x})$ in vector notation.

As mentioned previously, the third- and higher-order densities are very cumbersome to write out explicitly; therefore, we will examine the bivariate density in some detail in order to gain insight into the general multivariate normal density function. A sketch of $f_{X_1 X_2}(x_1, x_2)$ is shown in Fig. 1.23. The bivariate normal density is seen to be a smooth hill-shaped surface over the x_1, x_2 plane with the peak of the hill occurring directly above the point (m_1, m_2) in the x_1, x_2 plane. Equal-height contours on the $f_{X_1 X_2}$ surface project into ellipses in the x_1, x_2 plane, and the one shown in Fig. 1.23 is typical for a positive correlation coefficient ρ . Points on the ellipse may be thought of as equally likely combinations of x_1 and x_2 . If $\rho = 0$, we have the case where X_1 and X_2 are uncorrelated, and the ellipses have their semimajor and semiminor axes parallel

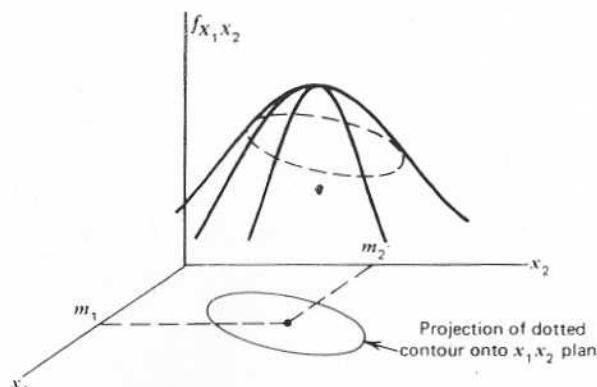


Figure 1.23. Bivariate normal density function.

to the x_1 and x_2 axes. If we specialize further and let $\sigma_1 = \sigma_2$ (and $\rho = 0$), the ellipses degenerate to circles. In the other extreme, as $|\rho|$ approaches unity, the ellipses become more and more eccentric.

The uncorrelated case where $\rho = 0$ is of special interest, and in this case f_{X_1, X_2} reduces to the form given in Eq. (1.15.11).

For uncorrelated X_1 and X_2 :

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[\frac{(x_1 - m_1)^2}{\sigma_1^2} + \frac{(x_2 - m_2)^2}{\sigma_2^2} \right] \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1 - m_1)^2/2\sigma_1^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x_2 - m_2)^2/2\sigma_2^2} \end{aligned} \quad (1.15.11)$$

The joint density function is seen to factor into the product of $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. Thus, *two normal random variables that are uncorrelated are also statistically independent*. It is easily verified from Eq. (1.15.5) that this is also true for any number of uncorrelated normal random variables. This is exceptional, because in general zero correlation does not imply statistical independence. It does, however, in the Gaussian case.

Now try to visualize the three-variate normal density function. The locus of constant $f_{X_1, X_2, X_3}(x_1, x_2, x_3)$ will be a closed elliptically shaped surface with three axes of symmetry. These axes will be aligned with the x_1, x_2, x_3 axes for the case of zero correlation among the three variates. If, in addition to zero correlation, the variates have equal variances, the surface becomes spherical.

If we try to extend the geometric picture beyond the three-variate case, we run out of Euclidean dimensions. However, conceptually, we can still envision equal-likelihood surfaces in hyperspace, but there is no way of sketching a picture of such surfaces. Some general properties of multivariate normal random variables will be explored further in the next section.

1.16 LINEAR TRANSFORMATION AND GENERAL PROPERTIES OF NORMAL RANDOM VARIABLES

The general linear transformation of one set of normal random variables to another is of special interest in noise analysis. This will now be examined in detail.

We have just seen that the density function for jointly normal random variables X_1, X_2, \dots, X_n can be written in matrix form as

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}_X|^{1/2}} \exp \left\{ -\frac{1}{2} [(\mathbf{x} - \mathbf{m}_X)^T \mathbf{C}_X^{-1} (\mathbf{x} - \mathbf{m}_X)] \right\} \quad (1.16.1)$$

We have added the subscript X to \mathbf{m} and \mathbf{C} to indicate that these are the mean and covariance matrices associated with the \mathbf{X} random variable. We now define a new set of random variables Y_1, Y_2, \dots, Y_n that are linearly related to X_1, X_2, \dots, X_n via the equation

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (1.16.2)$$

where lowercase \mathbf{x} and \mathbf{y} indicate realizations of \mathbf{X} and \mathbf{Y} , \mathbf{b} is a constant vector, and \mathbf{A} is a square matrix that will be assumed to be nonsingular (i.e., invertible). We wish to find the density function for \mathbf{Y} , and we can use the methods of Section 1.14 to do so. In particular, the transformation is one-to-one; therefore, a generalized version of Eq. (1.14.32) may be used.

$$f_Y(\mathbf{y}) = f_X(\mathbf{x}(\mathbf{y})) |J\left(\frac{\mathbf{x}}{\mathbf{y}}\right)| \quad (1.16.3)$$

We must first solve Eq. (1.16.2) for \mathbf{x} in terms of \mathbf{y} . The result is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b} \quad (1.16.4)$$

Let the individual terms of \mathbf{A}^{-1} be denoted as

$$\mathbf{A}^{-1} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & \cdots & \cdots & d_{nn} \end{bmatrix} \quad (1.16.5)$$

The scalar equations represented by Eq. (1.16.4) are then

$$\begin{aligned} x_1 &= (d_{11}y_1 + d_{12}y_2 + \cdots) - (d_{11}b_1 + d_{12}b_2 + \cdots) \\ x_2 &= (d_{21}y_1 + d_{22}y_2 + \cdots) - (d_{21}b_1 + d_{22}b_2 + \cdots) \\ x_3 &= \cdots \text{etc.} \end{aligned} \quad (1.16.6)$$

The Jacobian of the transformation is then

$$J\left(\frac{\mathbf{x}}{\mathbf{y}}\right) = J\left(\frac{x_1, x_2, \dots}{y_1, y_2, \dots}\right) = \text{Det} \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \dots \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \dots \\ \dots & \dots & \dots \end{bmatrix}$$

$$= \text{Det} \begin{bmatrix} d_{11} & d_{21} & \dots \\ d_{12} & d_{22} & \dots \\ \dots & \dots & \dots \end{bmatrix} = \text{Det } (\mathbf{A}^{-1})^T = \text{Det } (\mathbf{A}^{-1}) \quad (1.16.7)$$

We can now substitute Eqs. (1.16.4) and (1.16.7) into Eq. (1.16.3). The result is

$$f_Y(\mathbf{y}) = \frac{|\text{Det } (\mathbf{A}^{-1})|}{(2\pi)^{n/2} |\mathbf{C}_X|^{1/2}} \exp \left\{ -\frac{1}{2} [(\mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b} - \mathbf{m}_X)^T \mathbf{C}_X^{-1} (\mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b} - \mathbf{m}_X)] \right\} \quad (1.16.8)$$

We find the mean of \mathbf{Y} by taking the expectation of both sides of the linear transformation

$$\mathbf{Y} = \mathbf{AX} + \mathbf{b}$$

Thus

$$\mathbf{m}_Y = \mathbf{Am}_X + \mathbf{b} \quad (1.16.9)$$

The exponent in Eq. (1.16.8) may now be written as

$$\begin{aligned} -\frac{1}{2} [(\mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{Am}_X)^T \mathbf{C}_X^{-1} (\mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{Am}_X)] \\ = -\frac{1}{2} [(\mathbf{y} - \mathbf{m}_Y)^T (\mathbf{A}^{-1})^T \mathbf{C}_X^{-1} \mathbf{A}^{-1} (\mathbf{y} - \mathbf{m}_Y)] \\ = -\frac{1}{2} [(\mathbf{y} - \mathbf{m}_Y)^T (\mathbf{AC}_X \mathbf{A}^T)^{-1} (\mathbf{y} - \mathbf{m}_Y)] \end{aligned} \quad (1.16.10)$$

The last step in Eq. (1.16.10) is accomplished by using the reversal rule for the inverse of triple products and noting that the order of the transpose and inverse operations may be interchanged. Also note that

$$|\text{Det } (\mathbf{A}^{-1})| = \frac{1}{|\text{Det } \mathbf{A}|} = \frac{1}{|\text{Det } \mathbf{A}|^{\frac{1}{2}} \cdot |\text{Det } \mathbf{A}^T|^{\frac{1}{2}}} \quad (1.16.11)$$

Substitution of the forms given in Eqs. (1.16.10) and (1.16.11) into Eq. (1.16.8) yields for f_Y

$$f_Y(\mathbf{y}) = \frac{1}{(2\pi)^{n/2} |\mathbf{AC}_X \mathbf{A}^T|^{1/2}} \exp \left\{ -\frac{1}{2} [(\mathbf{y} - \mathbf{m}_Y)^T (\mathbf{AC}_X \mathbf{A}^T)^{-1} (\mathbf{y} - \mathbf{m}_Y)] \right\} \quad (1.16.12)$$

It is apparent now that f_Y is also normal in form with the mean and covariance matrix given by

$$\mathbf{m}_Y = \mathbf{Am}_X + \mathbf{b} \quad (1.16.13)$$

and

$$\mathbf{C}_Y = \mathbf{AC}_X \mathbf{A}^T \quad (1.16.14)$$

Thus we see that *normality is preserved in a linear transformation*. All that is changed is the mean and the covariance matrix; the *form* of the density function remains unchanged.

There are, of course, an infinite number of linear transformations one can make on a set of normal random variables. The particular transformation, say \mathbf{S} , that produces a new covariance matrix $\mathbf{SC}_X \mathbf{S}^T$ that is diagonal is of special interest. This transformation yields a new set of normal random variables that are uncorrelated, and thus they are also statistically independent. In a given problem, we may not choose to actually make this change of variables, but it is important just to know that the variables can be decoupled and under what circumstances this can be done. It works out that a diagonalizing transformation will always exist if \mathbf{C}_X is positive definite.* In the case of a covariance matrix, this is the equivalent of saying that all the correlation coefficients are less than unity in magnitude. This will be demonstrated for the bivariate case, and the extension to higher-order cases is fairly obvious.

A symmetric matrix \mathbf{C} is said to be positive definite if the scalar $\mathbf{x}^T \mathbf{Cx}$ is positive for all nontrivial \mathbf{x} , that is, $\mathbf{x} \neq 0$. Writing out $\mathbf{x}^T \mathbf{Cx}$ explicitly for the 2×2 case yields

$$[x_1 x_2] \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c_{11}x_1^2 + 2c_{12}x_1x_2 + c_{22}x_2^2 \quad (1.16.15)$$

But, if \mathbf{C} is a covariance matrix,

$$c_{11} = \sigma_1^2, \quad c_{12} = \rho\sigma_1\sigma_2, \quad c_{22} = \sigma_2^2 \quad (1.16.16)$$

* The diagonalizing transformation is sometimes called the *similarity transformation*. In the situation at hand, \mathbf{C}_X is a symmetric, positive-definite matrix, and there will exist a set of linearly independent, orthogonal eigenvectors for such a matrix (7). The transforming matrix \mathbf{S} has as its columns the normalized eigenvectors of \mathbf{C}_X . In this very special case, $\mathbf{S}^{-1} = \mathbf{S}^T$.

Therefore, $\mathbf{x}^T \mathbf{C} \mathbf{x}$ is

$$\mathbf{x}^T \mathbf{C} \mathbf{x} = (\sigma_1 x_1)^2 + 2\rho(\sigma_1 x_1)(\sigma_2 x_2) + (\sigma_2 x_2)^2 \quad (1.16.17)$$

Equation (1.16.17) now has a simple geometric interpretation. Assume $|\rho| < 1$; ρ can then be related to the negative cosine of some angle θ , where $0 < \theta < \pi$. Equation (1.16.17) will then be recognized as the equation for the square of the "opposite side" of a general triangle; and this, of course, must be positive. Thus, a 2×2 covariance matrix is positive definite, provided $|\rho| < 1$.

It is appropriate now to summarize some of the important properties of multivariate normal random variables:

1. The probability density function describing a vector random variable \mathbf{X} is completely defined by specifying the mean and covariance matrix of \mathbf{X} .
2. The covariance matrix of \mathbf{X} is positive definite if the magnitudes of all correlation coefficients are less than unity.
3. If normal random variables are uncorrelated, they are also statistically independent.
4. A linear transformation of normal random variables leads to another set of normal random variables. A decoupling (decorrelating) transformation will always exist if the original covariance matrix is positive definite.
5. If the joint density function for n random variables is normal in form, all marginal and conditional densities associated with the n variates will also be normal in form. (This was not shown; see Problem 1.21.)

1.17 LIMITS, CONVERGENCE, AND UNBIASED ESTIMATORS

No discussion of probability could be complete without at least some mention of limits and convergence. To put this in perspective, we first review the usual deterministic concept of convergence. As an example, recall that the Maclaurin series for e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \quad (1.17.1)$$

This series converges uniformly to e^x for all real x in any finite interval. By convergence we mean that if a given accuracy figure is specified, we can find an appropriate number of terms such that the specified accuracy is met by a truncated version of the series. In particular, note that once we have determined how many terms are needed in the truncated series, this same number is good for all x within the interval, and there is nothing "chancy" about it. In contrast, we will see presently that such "100 percent sure" statements cannot

be made in probabilistic situations. A look at the sample mean of n random variables will serve to illustrate this.

Let X_1, X_2, \dots, X_n be independent random variables with identical probability density functions $f_X(x)$. In terms of an experiment, these may be thought of as ordered samples of the random variable X . Next, consider a sequence of random variables defined as follows:

$$\begin{aligned} Y_1 &= X_1 \\ Y_2 &= \frac{X_1 + X_2}{2} \\ Y_3 &= \frac{X_1 + X_2 + X_3}{3} \\ &\vdots \\ Y_n &= \frac{X_1 + X_2 + \dots + X_n}{n} \end{aligned} \quad (1.17.2)$$

The random variable Y_n is, of course, just the sample mean of the random variable X . We certainly expect Y_n to get closer to $E(X)$ as n becomes large. But, closer in what sense? This is the crucial question. It should be clear that any particular experiment could produce an "unusual" event in which the sample mean would differ from $E(X)$ considerably. On the other hand, quite by chance, a similar experiment might yield a sample mean that was quite close to $E(X)$. Thus, in this probabilistic situation, we cannot expect to find a fixed number of samples n that will meet a specified accuracy figure for all experiments. No matter how large we make n , there is always some nonzero probability that the very unusual thing will happen, and a particular experiment will yield a sample mean that is outside the specified accuracy. Thus, we can only hope for convergence in some sort of average sense and not in an absolute (100 percent sure) sense.

Let us now be more specific in this example, and let X (and thus X_1, X_2, \dots, X_n) be normal with mean m_X and variance σ_X^2 . From Section 1.16 we also know that the sample mean Y_n is a normal random variable. Since a normal random variable is characterized by its mean and variance, we now examine these parameters for Y_n . The expectation of a sum of elements is the sum of the expectations of the elements. Thus

$$\begin{aligned} E(Y_n) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \frac{1}{n} [E(X_1) + E(X_2) + \dots] \\ &= \frac{1}{n} [nE(X)] = m_X \end{aligned} \quad (1.17.3)$$

The sample mean is, of course, an estimate of the true mean of X , and we see from Eq. (1.17.3) that it at least yields $E(X)$ "on the average." Estimators that have this property are said to be unbiased. That is, an estimator is said to be *unbiased* if

$$E(\text{Estimate of } X) = E(X) \quad (1.17.4)$$

Consider next the variance of Y_n . Using Eq. (1.17.3) and recalling that the expectation of the sum is the sum of the expectations,

$$\begin{aligned} \text{Var } Y_n &= E[Y_n - E(Y_n)]^2 \\ &= E(Y_n^2 - 2Y_n m_X + m_X^2) \\ &= E(Y_n^2) - m_X^2 \end{aligned} \quad (1.17.5)$$

The sample mean Y_n may now be replaced with $\frac{1}{n}(X_1 + X_2 + \dots + X_m)$; and, after squaring and some algebraic simplification, Eq. (1.17.5) reduces to

$$\begin{aligned} \text{Var } Y_n &= \frac{1}{n} \text{Var } X \\ &= \frac{\sigma_X^2}{n} \end{aligned} \quad (1.17.6)$$

Thus we see that the variance of the sample mean decreases with increasing n and eventually goes to zero as $n \rightarrow \infty$.

The probability density functions associated with the sample mean are shown in Fig. 1.24 for three values of n . It should be clear from the figure that convergence of some sort takes place as $n \rightarrow \infty$. However, no matter how large we make n , there will still remain a nonzero probability that Y_n will fall outside some specified accuracy interval. Thus, we only have convergence in a statistical sense and not in an absolute deterministic sense.

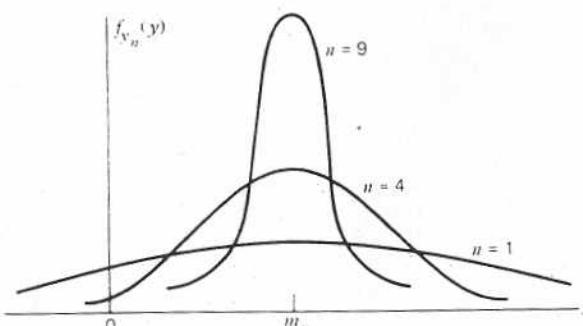


Figure 1.24. Probability density functions illustrating convergence of the sample mean.

There are a number of types of statistical convergence that have been defined and are in common usage (8). We look briefly at two of these. Consider a sequence of random variables Y_1, Y_2, \dots, Y_n . The sequence Y_n is said to converge in the mean to Y if

$$\lim_{n \rightarrow \infty} E[(Y_n - Y)^2] = 0 \quad (1.17.7)$$

Convergence in the mean is sometimes abbreviated as

$$\text{l.i.m. } Y_n = Y \quad (1.17.8)$$

where l.i.m. denotes "limit in the mean."

The sequence Y_n converges in probability to Y if

$$\lim_{n \rightarrow \infty} P(|Y_n - Y| \geq \epsilon) = 0 \quad (1.17.9)$$

where ϵ is an arbitrarily small positive number.

It should be clear from Eqs. (1.17.3) and (1.17.6) that the sample mean converges "in the mean" to m_X . It also converges in probability because the area under the "tails" of the probability density outside a specified interval about m_X goes to zero as $n \rightarrow \infty$. Roughly speaking, convergence in the mean indicates that the dispersion (variance) about the limiting value shrinks to zero in the limit. Similarly, convergence in probability means than an arbitrarily small accuracy criterion is met with a probability of one as $n \rightarrow \infty$. Davenport and Root (8) point out that convergence in the mean is a more severe requirement than convergence in probability. Thus, if a sequence converges in the mean, we are also assured that it will converge in probability. The converse is not true though, because convergence in probability is a "looser" sort of criterion than convergence in the mean.

Problems

- 1.1. In straight poker, five cards are dealt to each player from a deck of ordinary playing cards. What is the probability that a player will be dealt a flush (i.e., five cards all of one suit)?
- 1.2. In the game of blackjack the player is initially dealt two cards from a deck of ordinary playing cards. Without going into all the details of the game, it will suffice to say here that the best possible hand one could receive on the initial deal is a combination of an ace of any suit and any face card or 10. What is the probability that the player will be dealt this combination?
- 1.3. An urn contains two white, one black, and four red balls. Three balls are drawn out simultaneously.
 - (a) How many elements are in the sample space for this experiment?
 - (b) What is the probability that the draw will produce one white and two red balls?

1.4. Five dice are rolled simultaneously.

- (a) Describe the sample space for this experiment.
- (b) What is the probability that exactly one 6 will be rolled?
- (c) What is the probability that at least one 6 will be rolled?

1.5. Consider a horse race with five entries. The approximate win odds posted by the track just prior to the race are:

Horse Number	Approximate Odds
1	2 to 1
2	3 to 1
3	5 to 1
4	5 to 1
5	11 to 1

Track "odds" simply refer to additional amount returned to the bettor in event of a win. For example, if the bettor places a bet on horse #1 and that horse wins with 2 to 1 odds, the track will return 1 (the original wager) plus 2 (for winning), or 3 dollars for each dollar wagered.

Assume the race is a chance event and let the sample space consist of five elements: Horse #1 wins, Horse #2 wins, etc. Ignore the small percentage kept by the track (the odds are approximate anyway), and make a reasonable probability assignment to each element in the sample space for this situation.

1.6. Sketch a Venn diagram for three events A , B , and C . Use the diagram to justify the following relationship:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) \\ &\quad - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C) \end{aligned} \quad (\text{P1.6})$$

1.7. Contract bridge is played with an ordinary deck of 52 playing cards. There are four players with players on opposite sides of the table being partners. One player acts as dealer and deals each player 13 cards. A bidding sequence then takes place, and this establishes the trump suit and names the player who is to attempt to take a certain number of tricks. This player is called the declarer. The play begins with the player to the declarer's left leading a card. The declarer's partner's hand is then laid out on the table face up for everyone to see. This then enables the declarer to see a total of 27 cards as the play begins. Knowledge of these cards will, of course, affect his or her strategy in the subsequent play.

Suppose the declarer sees 11 of the 13 trump cards as the play begins. We will assume that the opening lead was not a trump, which leaves 2 trumps outstanding in the opponent's hands. The disposition of these is, of course, unknown to the declarer. There are, however, a limited number of possibilities:

- (a) Both trumps lie with the opponent to the left and none to the right.
- (b) Both trumps are to the right and none to the left.
- (c) The two trumps are split, one in each opponent's hand.

Compute the probabilities for each of the (a), (b), and (c) possibilities. (*Hint:* Rather than look at all possible combinations, look at numbers of combinations for 25 specific cards held by the opponents just after the opening lead. Two of these will, of course, be

specific trump cards. The resulting probability will be the same regardless of the particular choice of specific cards.)

1.8. In the game of craps the casino also has ways of making money other than betting against the player rolling the dice. The other participants standing around the table may place side bets with the casino while waiting their turn with the dice. One such side bet at some tables is "Don't Pass—Bar 12." The participant placing such a side bet wins if the player rolling the dice loses, except for the case of a 12 on the first roll of the dice, and in that case the result is a standoff. In other words, the side bettor is betting against the roller, except that the rules are changed slightly in the side-bet wager to make it such that 12 on the first roll does not win for either the casino or the side bettor. Presumably, this tips the odds in favor of the casino for the side bet.

- (a) Find the probability of winning when placing a side bet "Don't Pass—Bar 12."
- (b) It is often said that the casinos make a higher percentage (on the average) on side bets than they do on the regular bet with the player rolling the dice. Compare the results of this problem with those of Example 1.2, relative to the average percentage "take" for the casino. [It works out that this side bet is fairly even. Others are not so favorable. See Goren (11).]

1.9. Assume equal likelihood for the birth of boys and girls. What is the probability that a four-child family chosen at random will have two boys and two girls, irrespective of the order of birth? (Note that the answer is not $\frac{1}{2}$ as might be suspected at first glance.)

1.10. Consider a sequence of random binary digits, zeros and ones. Each digit may be thought of as an independent sample from a sample space containing two elements, each having a probability of $\frac{1}{2}$. For a six-digit sequence, what is the probability of having:

- (a) Exactly 3 zeros and 3 ones arranged in any order?
- (b) Exactly 4 zeros and 2 ones arranged in any order?
- (c) Exactly 5 zeros and 1 one arranged in any order?
- (d) Exactly 6 zeros?

1.11 A certain binary message is n bits in length. If the probability of making an error in the transmission of a single bit is p , and if the error probability does not depend on the outcome of any previous transmissions, show that the probability of occurrence of exactly k bit errors in a message is

$$P(k \text{ errors}) = \binom{n}{k} p^k (1-p)^{n-k} \quad (\text{P1.11})$$

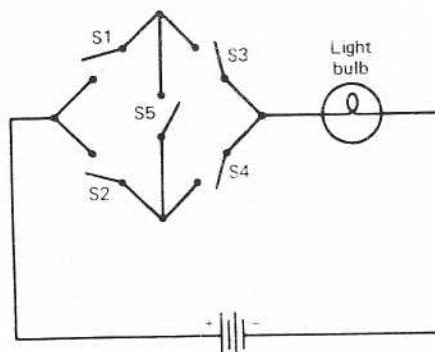
The quantity $\binom{n}{k}$ denotes the number of combinations of n things taken k at a time. (This is a generalization of Problems 1.9 and 1.10.)

1.12 There are 5 switches labeled S_1, S_2, \dots, S_5 in the circuit shown. Assume that the position of each switch is governed by chance and that each is equally likely to be open or closed.

- (a) What is the probability that the light bulb is "on"?
- (b) Given that the light bulb is "on," what is the probability that S_5 is closed?

1.13. The random variable X may take on all values between 0 and 2, with all values within this range being equally likely.

- (a) Sketch the probability density function for X .
- (b) Sketch the cumulative probability distribution function for X .
- (c) Calculate $E(X)$, $E(X^2)$, and $\text{Var } X$.



Problem 1.12

1.14. A random variable X has a probability density function as shown below.

- Sketch the cumulative distribution function for X .
- What is the variance of X ?

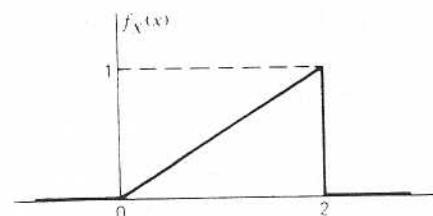
1.15. Resistors are a common component in nearly all electronic equipment. Suppose a resistor company manufactures a large lot of resistors whose nominal value is to be 100Ω . Because of manufacturing tolerances this results in resistance values spread between 90 and 110Ω (± 10 percent of nominal). For the sake of simplicity let us assume the spread is uniform between the two extremes. Let us further speculate that the manufacturer sorts out all the ± 5 percent tolerance resistors for sale at a premium price; the remaining resistors are then classified as "10 percent resistors" and are to be sold at the regular price. Let the resistance value of such a "10 percent resistor" be a random variable X .

- What is the probability density function for X ?
- What is the probability that X will lie within 8 percent of the nominal value?

1.16. Suppose the resistance values of Problem 1.15 were normally distributed with a mean of 100Ω and a standard deviation of 5Ω . What is the probability that a sample resistance out of the batch will lie between 90 and 110Ω ? (Note: Negative resistance is not physically possible, so a normal distribution can only be an approximation in this situation.)

1.17. A random variable X whose probability density function is given by

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



Problem 1.14

is said to have an exponential probability density function. This density function is sometimes used to describe the failure of equipment components (9.10). That is, the probability that a particular component will fail within time T is

$$P(\text{failure}) = \int_0^T f_X(x) dx \quad (\text{P1.17})$$

Note α is a parameter that may be adjusted to fit the situation at hand.

- Find α for an electronic component whose average lifetime is 10,000 hours. ("Average" is used synonymously with "expectation" here.)
- Suppose we wish the probability of failure for the component of part (a) to be less than .01, that is, we wish the reliability to be .99. For what time span might we expect a reliability of .99?

1.18. Consider a sack containing several identical coins whose sides are labeled +1 and -1. A certain number of coins are withdrawn and tossed simultaneously. The algebraic sum of the numbers resulting from the toss is a discrete random variable. Sketch the probability density function associated with the random variable for the following situations:

- One coin is tossed.
- Two coins are tossed.
- Five coins are tossed.
- Ten coins are tossed.

The density functions in this case consist of impulses. In the sketches, represent impulses with "arrows" whose lengths are proportional to the magnitudes of the impulses. (This is the discrete analog of Example 1.21. Note the tendency toward central distribution.)

1.19. Two players are matching pennies. Each player begins with a stack of pennies and they uncover their pennies one at a time. By mutual agreement, pennies that match are taken by one player and those that do not match are taken by the other. When either player's original stack is exhausted, the players restack their available pennies and repeat the matching process. This is continued until one or the other player has won all the pennies.

Suppose player A has two pennies and player B has one. Sketch the probability density function for the number of trials required to end the game. The game ends when either player runs out of pennies. (Just as in Problem 1.18, use arrows to represent impulses in the sketch.)

1.20. A player throws two dice; one is red and the other white. After the dice are thrown, an observer can see that the red die turned up 5. The observer cannot see the white die.

- What is the probability, conditioned on the observation of the red die, that the player has thrown a sum of 7 or 11? (This would be an instantaneous win on the first throw in craps.)
- Compare the results of part (a) with the unconditional probability of winning on the first throw in craps.

1.21. Let the sum of the dots resulting from the throw of two dice be a discrete random variable X . The probabilities associated with their permissible values $2, 3, \dots, 12$ are easily found by itemizing all possible results of the throw (or from Table 1.2). Find $E(X)$ and $\text{Var } X$.

1.22. Discrete random variables X and Y may each take on integer values 1, 3, and 5,

and the joint probability of X and Y is given in the table below.

$X \backslash Y$	1	3	5
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
3	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
5	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

- (a) Are random variables X and Y independent?
- (b) Find the unconditional probability $P(Y = 5)$.
- (c) What is the conditional probability $P(Y = 5|X = 3)$?

1.23. The diagram shown below gives the error characteristics of a hypothetical binary transmission system. The numbers shown next to the arrows are the conditional probabilities of Y given X . The unconditional probabilities for X are shown to the left of the figure. Find:

- (a) The conditional probabilities $P(X = 0|Y = 1)$ and $P(X = 0|Y = 0)$.
- (b) The unconditional probabilities $P(Y = 0)$ and $P(Y = 1)$.
- (c) The joint probability array for $P(X, Y)$.

1.24. The Poisson distribution is defined as

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (\text{P1.24})$$

where λ is a positive real parameter and $k = 0, 1, 2, \dots$. This discrete distribution appears in a wide variety of applications in which events occur at random with time (5.9). For example, a particular telephone call handled by an office might occur with equal likelihood at any time during the working day. Let the average number of calls handled per unit time be α . Then the probability of exactly k calls occurring in a time interval T is

$$P(k) = \frac{(\alpha T)^k}{k!} e^{-\alpha T}$$

Obviously, in this instance αT is the λ parameter of the Poisson distribution.

- (a) Let αT be 2 and sketch the corresponding probability density as a function of k for $k = 0, 1, 2, \dots, 10$. (As before, use arrows for impulses.)
- (b) Find the mean for the Poisson distribution. (Hint: The characteristic function discussed in Section 1.8 is useful here. Answer: λ .)

Total Probability $P(X)$	Transmitter X	Receiver Y
$P(X = 0) = .75$	0	.9
$P(X = 1) = .25$	1	.2 .8

Problem 1.23

1.25. A large power system has an average of three outages per year somewhere in the system. If the outages are assumed to occur randomly, what is the probability of going without an outage for the next year? (See Problem 1.24 for a discussion of the Poisson distribution.)

1.26. Records show that a certain city has had an average of 1 earthquake per decade when averaged over the past 200 years. Assume that earthquakes occur at random and find the probability that at least one earthquake will occur within:

- (a) The next decade.
- (b) A person's normal lifespan (about 70 years).

(See Problem 1.24 for a discussion of the Poisson distribution.)

1.27. The Rayleigh probability density function is defined as

$$f_R(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2} \quad (\text{P1.27})$$

where σ^2 is a parameter of the distribution (see Example 1.24).

- (a) Find the mean and variance of a Rayleigh distributed random variable R .
- (b) Find the mode of R (i.e., the most likely value of R).

1.28. Three similar "unfair" coins are tossed simultaneously. The coins are unfair in that $P(\text{Heads}) = .6$ and $P(\text{Tails}) = .4$. Let the discrete random variable X be defined to be the number of heads that results from the toss of the three coins. Find the discrete probability distribution associated with the random variable X .

1.29. The target shooting example of Section 1.14 led to the Rayleigh density function specified by Eq. (1.14.35) (also repeated in Problem 1.27).

- (a) Show that the probability that a hit will lie within a specified distance R_0 from the origin is given by

$$P(\text{Hit lies within } R_0) = 1 - e^{-R_0^2/2\sigma^2} \quad (\text{P1.29})$$

- (b) The value of R_0 in Eq. (P1.29) that yields a probability of .5 is known as the circular probable error (or circular error probable, CEP). Find the CEP in terms of σ .

1.30. Find the mean and variance of the output of a half-wave rectifier driven by Gaussian noise. (The probability density function for the output is given in Example 1.18.)

1.31. Consider a random variable X with an exponential probability density function

$$f_X(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Find:

- (a) $P(X \geq 2)$
- (b) $P(1 \leq X \leq 2)$
- (c) $E(X)$, $E(X^2)$, and $\text{Var } X$.

1.32. Random variables X and Y have a joint probability density function defined as follows:

$$f_{XY}(x, y) = \begin{cases} .25, & -1 \leq x \leq 1 \text{ and } -1 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Are random variables X and Y statistically independent? (Hint: Integrate with respect to

appropriate dummy variables to obtain $f_X(x)$ and $f_Y(y)$. Then check to see if the product of f_X and f_Y is equal to f_{XY} .)

1.33. Random variables X and Y have a joint probability density function

$$f_{XY}(x, y) = \begin{cases} e^{-(x+y)}, & x \geq 0 \text{ and } y \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Find:

- (a) $P(X \leq \frac{1}{2})$
- (b) $P[(X + Y) \leq 1]$
- (c) $P[(X \text{ or } Y) \geq 1]$
- (d) $P[(X \text{ and } Y) \geq 1]$

1.34. Are the random variables of Problem 1.33 statistically independent?

1.35. Random variables X and Y are statistically independent and their respective probability density functions are

$$f_X(x) = \frac{1}{2}e^{-|x|}$$

$$f_Y(y) = e^{-2|y|}$$

Find the probability density function associated with $X + Y$. (*Hint:* Fourier transforms are helpful here.)

1.36. Random variable X has a probability density function

$$f_X(x) = \begin{cases} \frac{1}{2}, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Random variable Y is related to X through the equation

$$y = x^3 + 1$$

What is the probability density function for Y ?

1.37. X and Y are independent, zero-mean random variables with variances σ_X^2 and σ_Y^2 . Another set of random variables U and V are related to X and Y through the equations

$$u = 2x + y$$

$$v = x - y$$

Find the correlation coefficient of U and V . Let $\sigma_X^2 = \sigma_Y^2$.

1.38. The vector Gaussian random variable

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is completely described by its mean and covariance matrix. In this case they are

$$\mathbf{m}_X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\mathbf{C}_X = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}$$

Now consider another vector random variable \mathbf{Y} that is related to \mathbf{X} by the equation

$$\mathbf{y} = \mathbf{Ax} + \mathbf{b}$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Find the mean and covariance matrix for \mathbf{Y} .

1.39. The general bivariate normal density function is given by Eq. (1.16.1). It should be apparent that the unconditional density functions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are also normal in form. (A linear transformation yielding uncorrelated random variables may be made, and then the normal form is obvious.) Show that the conditional density functions $f_{X_1|X_2}(x_1|x_2)$ and $f_{X_2|X_1}(x_2|x_1)$ are also normal in form. (*Hint:* Simply noting the quadratic form in the exponential is sufficient justification of normality.)

1.40. A pair of random variables, X and Y , have a joint probability density function

$$f_{XY}(x, y) = \begin{cases} 1, & 0 \leq y \leq 2x \text{ and } 0 \leq x \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find:

$$E(X|Y = .5)$$

[*Hint:* Use Eq. (1.11.19) to find $f_{X|Y}(x)$ for $y = .5$, and then integrate $xf_{X|Y}(x)$ to find $E(X|Y = .5)$.]

References Cited in Chapter 1

1. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Cambridge, Mass.: MIT Press; and New York: Wiley, 1949.
2. S. O. Rice, "Mathematical Analysis of Noise," *Bell System Tech. J.*, 23, 282–332 (1944); 24, 46–256 (1945).
3. A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, 3rd ed., New York: McGraw-Hill, 1974.
4. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd ed., New York: Wiley, 1957.
5. P. Beckmann, *Probability in Communication Engineering*, New York: Harcourt, Brace, and World, 1967.
6. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965.
7. S. I. Grossman, *Elementary Linear Algebra*, Florence, KY: Wadsworth, Inc., 1980.
8. W. B. Davenport, Jr., and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.
9. I. F. Blake, *An Introduction to Applied Probability*, New York: Wiley, 1979.
10. H. J. Larson and B. O. Shubert, *Probabilistic Models in Engineering Sciences (Vol. 1)*, New York: Wiley, 1979.
11. C. H. Goren, *Go With the Odds*, New York: Macmillan, 1969.

Additional References on Probability

12. P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles*, New York: McGraw-Hill, 1980.
13. G. R. Cooper and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, New York: Holt, Rinehart, and Winston, 1971.
14. A. M. Breipohl, *Probabilistic Systems Analysis*, New York: Wiley, 1970.
15. J. L. Melsa and A. P. Sage, *An Introduction to Probability and Stochastic Processes*, Englewood Cliffs, N.J.: Prentice-Hall, 1973.

CHAPTER 2**Mathematical Description
of Random Signals**

The concept of frequency spectrum is familiar from elementary physics, and so it might seem appropriate to begin our discussion of noiselike signals with their spectral description. This approach, while intuitively appealing, leads to all sorts of difficulties. The only really careful way to describe noise is to begin with a probabilistic description and then proceed to derive the associated spectral characteristics from the probabilistic model. We now proceed toward this end.

2.1 CONCEPT OF A RANDOM PROCESS

We should begin by distinguishing between deterministic and random signals. Usually, the signals being considered here will represent some physical quantity such as voltage, current, distance, temperature, and so forth. Thus, they are real variables. Also, time will usually be the independent variable, although this does not necessarily need to be the case. A signal is said to be *deterministic* if it is exactly predictable for the time span of interest. Examples would be

- (a) $x(t) = 10 \sin 2\pi t$ (Sine wave)
- (b) $x(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}$ (Unit step)
- (c) $x(t) = \begin{cases} 1 - e^{-t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$ (Exponential response)

Notice that there is nothing "chancy" about any of these signals. They are described by functions in the usual mathematical sense; that is, specify a numerical value of t and the corresponding value of x is determined. We are usually able to write the functional relationship between x and t explicitly. However, this is not really necessary. All that is needed is to know conceptually that a functional relationship exists.

In contrast with a deterministic signal, a *random signal* always has some element of chance associated with it. Thus it is not predictable in a deterministic sense. Examples of random signals are:

- (d) $X(t) = 10 \sin (2\pi t + \theta)$, where θ is a random variable uniformly distributed between 0 and 2π .

- (e) $X(t) = A \sin(2\pi t + \theta)$, where θ and A are independent random variables with known distributions.
 (f) $X(t)$ = A noiselike signal with no particular deterministic structure—one that just wanders on aimlessly ad infinitum.

Since all of these signals have some element of chance associated with them, they are random signals. Signals such as (d), (e), and (f) are formally known as *random* or *stochastic processes*, and we will use the terms *random* and *stochastic* interchangeably throughout the remainder of the book.*

Let us now consider the description of signal (f) in more detail. It might be the common audible radio noise that was mentioned in Chapter 1. If we looked at an oscillographic recording of the radio speaker current, it might appear as shown in Fig. 2.1. We might expect such a signal to have some kind of spectral description, because the signal is audible to the human ear. Yet the precise mathematical description of such a signal is remarkably elusive, and it eluded investigators prior to the 1940s (3,4).

Imagine sampling the noise shown in Fig. 2.1 at a particular point in time, say t_1 . The numerical value obtained would be governed largely by chance, which suggests it might be considered to be a random variable. However, with random variables we must be able to visualize a conceptual statistical experiment in which samples of the random variable are obtained under identical chance circumstances. It would not be proper in this case to sample X by taking successive time samples of the same signal, because, if they were taken very close together, there would obviously be a close statistical connection among nearby samples. Therefore, the conceptual experiment in this case must consist of many “identical” radios, all playing simultaneously, all being tuned away from regular stations in different portions of the broadcast band, and all having their volumes turned up to the same sound level. This then leads to the notion of an ensemble of similar noiselike signals as shown in Fig. 2.2.

It can be seen then that a random process is a set of random variables that unfold with time in accordance with some conceptual chance experiment. Each of the noiselike time signals so generated is called a *sample realization* of the process. Samples of the individual signals at a particular time t_1 would then be sample realizations of the random variable $X(t_1)$. Four of these are illustrated in Fig. 2.2 as $X_A(t_1)$, $X_B(t_1)$, $X_C(t_1)$, and $X_D(t_1)$. If we were to sample at a different

*We need to recognize a notational problem here. Denoting the random process as $X(t)$ implies that there is a functional relationship between X and t . This, of course, is not the case because $X(t)$ is governed by chance. For this reason, some authors (1,2) prefer to use a subscript notation, that is, X_t rather than $X(t)$, to denote a random time signal. X_t then “looks” like a random variable with time as a parameter, which is precisely what it is. This notation, however, is not without its own problems. Suffice it to say, in most engineering literature, time random processes are denoted with “parentheses t ” rather than “subscript t .” We will do likewise, and the reader will simply have to remember that $X(t)$ does not mean function in the usual mathematical sense when $X(t)$ is a random process.

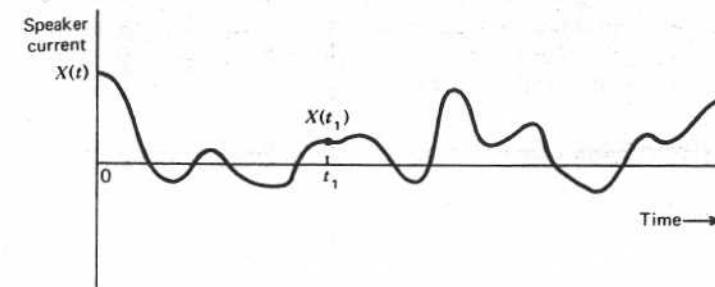


Figure 2.1. Typical radio noise signal.

time, say t_2 , we would obtain samples of a different random variable $X(t_2)$, and so forth. Thus, in this example, an infinite set of random variables is generated by the random process $X(t)$.

The radio experiment just described is an example of a *continuous random process* in that time evolves in a continuous manner. In this example, the probability density function describing the amplitude variation also happens to be continuous. However, random processes may also be discrete either in time or amplitude, as will be seen in the following two examples.

Example 2.1 Consider a card player with a deck of standard playing cards numbered from 1 (ace) through 13 (king). The deck is shuffled and the player picks a card at random and observes its number. The card is then replaced, the deck reshuffled, and another card is observed. This process is then repeated at

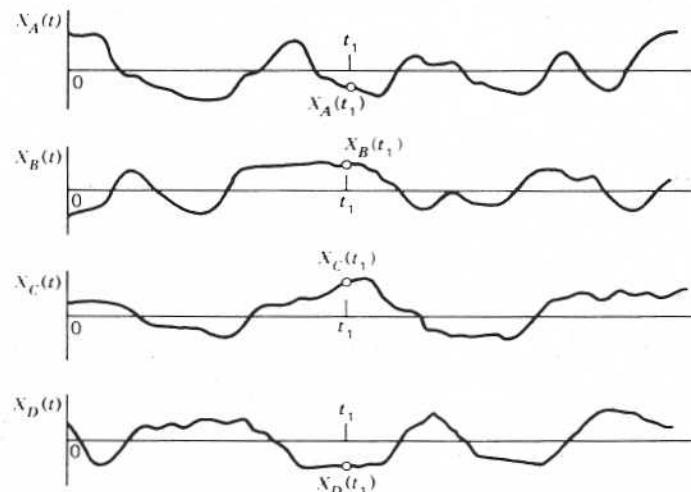


Figure 2.2. Ensemble of sample realizations of a random process.

unit intervals of time and continued on ad infinitum. The random process so generated would be discrete both in time and "amplitude," provided we say that the observed number only applies at the precise instant of time it is observed. ■

The preceding description would, of course, only generate one sample realization of the process. In order to obtain an ensemble of sample signals, we need to imagine an ensemble of card players, each having similar decks of cards and each generating a different (but statistically similar) sample realization of the process. ■

Example 2.2 Imagine a sack containing a large quantity of sample numbers taken from a zero-mean, unity-variance normal distribution. An observer reaches into the sack at unit intervals of time and observes a number with each trial. In order to avoid exact repetition, he does not replace the numbers during the experiment. This process would be discrete in time, as before, but continuous in amplitude. Also, the conceptual experiment leading to an ensemble of sample realizations of the process would involve many observers, each with a separate sack of random numbers. ■

2.2 PROBABILISTIC DESCRIPTION OF A RANDOM PROCESS

As mentioned previously, one can usually write out the functional form for a deterministic signal explicitly; for example, $s(t) = 10 \sin 2\pi t$, or $s(t) = t^2$, etc. No such deterministic description is possible for random signals because the numerical value of the signal at any particular time is governed by chance. Thus we should expect our description of noiselike signals to be somewhat vaguer than that for deterministic signals. One way to specify a random process is to describe in detail the conceptual chance experiment giving rise to the process. Examples 2.1 and 2.2 illustrate this way of describing a random process. The following two examples will illustrate this further.

Example 2.3 Consider a time signal (e.g., a voltage) that is generated according to the following rules: (a) The waveform is generated with a sample-and-hold arrangement where the "hold" interval is 1 sec; (b) the successive amplitudes are independent samples taken from a set of random numbers with uniform distribution from -1 to +1; and (c) the first switching time after $t = 0$ is a random variable with uniform distribution from 0 to 1. (This is equivalent to saying the time origin is chosen at random.) A typical sample realization of this process is shown in Fig. 2.3. Note the process mean is zero and its mean-square value works out to be one-third. [This is obtained from item (b) of the description.] ■

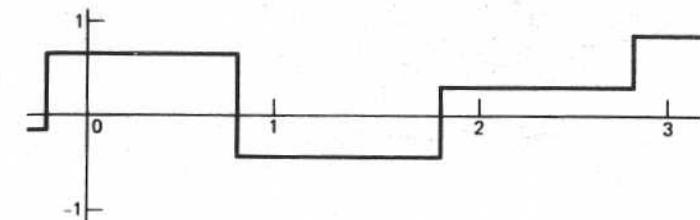


Figure 2.3. Sample signal for Example 2.3.

Example 2.4 Consider another time function generated with a sample-and-hold arrangement with these properties: (a) The "hold" interval is 0.2 sec, (b) the successive amplitudes are independent samples obtained from a zero-mean normal distribution with a variance of one-third, and (c) the switching points occur at multiples of .2 units of time; that is, the time origin is not chosen at random in this case. A sketch of a typical waveform for this process is shown in Fig. 2.4. ■

Now, from Examples 2.3 and 2.4 it should be apparent that if we simply say, "Noiselike waveform with zero mean and mean-square value of one-third," we really are not being very definite. Both processes of Examples 2.3 and 2.4 would satisfy these criteria, but yet they are quite different. Obviously, more information than just mean and variance is needed to completely describe a random process. We will now explore the "description" problem in more detail.

A more typical "noiselike" signal is shown in Fig. 2.5. The times indicated, t_1, t_2, \dots, t_k , have been arranged in ascending order, and the corresponding sample values X_1, X_2, \dots, X_k are, of course, random variables. Note that we have abbreviated the notation and have let $X(t_1) = X_1, X(t_2) = X_2, \dots$, etc. Obviously the first-order probability density functions $f_{X_1}(x), f_{X_2}(x), \dots, f_{X_k}(x)$ are important in describing the process because they tell us something about the process amplitude distribution. In Example 2.3, $f_{X_1}(x), f_{X_2}(x), \dots$,

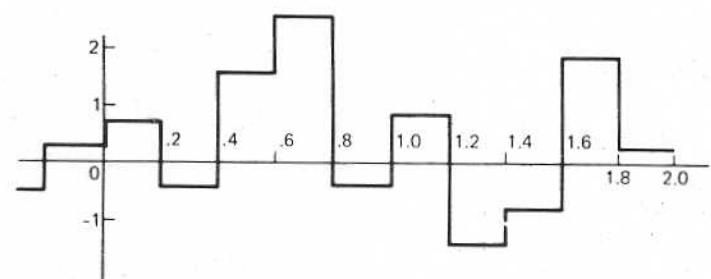


Figure 2.4. Typical waveform for Example 2.4.

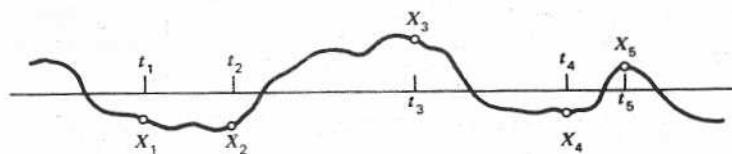


Figure 2.5. Sample signal of a typical noise process.

$f_{X_k}(x)$, are all identical density functions and are given by [using $f_{X_1}(x)$ as an example]

$$f_{X_1}(x) = \begin{cases} \frac{1}{2} & -1 \leq x \leq 1 \\ 0 & |x| > 1 \end{cases}$$

The density functions are not always identical for X_1, X_2, \dots, X_k ; they just happened to be in this simple example. In Example 2.4, the density functions describing the amplitude distribution of the X_1, X_2, \dots, X_k random variables are again all the same, but in this case they are normal in form with a variance of one-third. Note that the first-order densities tell us something about the relative distribution of the process amplitude as well as its mean and mean-square value.

It should be clear that the joint densities relating any pair of random variables, for example, $f_{X_1 X_2}(x_1, x_2), f_{X_1 X_3}(x_1, x_3)$, etc., are also important in our process description. It is these density functions that tell us something about how rapidly the signal changes with time, and these will eventually tell us something about the signal's spectral content. Continuing on, the third, fourth, and subsequent higher-order density functions provide even more detailed information about the process in probabilistic terms. However, this leads to a formidable description of the process, to say the least, because a k -variate density function is required where k can be any positive integer. Obviously, we will not usually be able to specify this k th order density function explicitly. Rather, this usually must be done more subtly by providing, with a word description or otherwise, enough information about the process to enable one to write out any desired higher-order density function; but the actual "writing it out" is usually not done.

Recall from probability theory that two random variables X and Y are said to be statistically independent if their joint density function can be written in product form

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad (2.2.1)$$

Similarly, random processes $X(t)$ and $Y(t)$ are statistically independent if the joint density for any combination of random variables of the two processes can be written in product form, that is, $X(t)$ and $Y(t)$ are independent if

$$f_{X_1 X_2 \dots Y_1 Y_2 \dots} = f_{X_1 X_2 \dots} f_{Y_1 Y_2 \dots} \quad (2.2.2)$$

In Eq. (2.2.2) we are using the shortened notation $X_1 = X(t_1), X_2 = X(t_2), \dots$, and $Y_1 = Y_1(t'_1), Y_2 = Y_2(t'_2), \dots$, where the sample times do not have to be the same for the two processes.

In summary, the test for completeness of the process description is this: Is enough information given to enable one, conceptually at least, to write out the k th order probability density function for any k ? If so, the description is as complete as can be expected; if not, it is incomplete to some extent, and radically different processes may fit the same incomplete description.

2.3 GAUSSIAN RANDOM PROCESS

There is one special situation where an explicit probability density description of the random process is both feasible and appropriate. This case is the *Gaussian* or *normal* random process. It is defined as one in which *all the density functions describing the process are normal in form*. Note that it is not sufficient that just the "amplitude" of the process be normally distributed; all higher-order density functions must also be normal! As an example, the process defined in Example 2.4 has a normal first-order density function, but closer scrutiny will reveal that its second-order density function is not normal in form. Thus the process is not a Gaussian process.

The multivariate normal density function was discussed in Section 1.15. It was pointed out there that matrix notation makes it possible to write out all k -variate density functions in the same compact matrix form, regardless of the size of k . All we have to do is specify the vector random-variable mean and covariance matrix, and the density function is specified. In the case of a Gaussian random process the "variates" are the random variables $X(t_1), X(t_2), \dots, X(t_k)$, where the points in time may be chosen arbitrarily. Thus enough information must be supplied to specify the mean and covariance matrix regardless of the choice of t_1, t_2, \dots, t_k . Examples showing how to do this will be deferred for the moment, because it is expedient first to introduce the basic ideas of stationarity and correlation functions.

2.4 STATIONARITY, ERGODICITY, AND CLASSIFICATION OF PROCESSES

A random process is said to be *time stationary* or simply *stationary* if the density functions describing the process are invariant under a translation of time. That is, if we consider a set of random variables $X_1 = X(t_1), X_2 = X(t_2), \dots, X_k = X(t_k)$, and also a translated set $X'_1 = X(t_1 + \tau), X'_2 = X(t_2 + \tau), \dots, X'_k = X(t_k + \tau)$, then

$X'_k = X(t_k + \tau)$, the density functions $f_{X_1}, f_{X_1 X_2}, \dots, f_{X_1 X_2 \dots X_k}$ describing the first set would be identical in form to those describing the translated set. Note that this applies to all the higher-order density functions. The adjective *strict* is also used occasionally with this type of stationarity to distinguish it from wide-sense stationarity which is a less restrictive form of stationarity. This will be discussed later in Section 2.5 on correlation functions.

A random process is said to be *ergodic* if time averaging is equivalent to ensemble averaging. In a qualitative sense this implies that a single sample time signal of the process contains all possible statistical variations of the process. Thus, no additional information is to be gained by observing an ensemble of sample signals over the information obtained from a one-sample signal, for example, one long strip-chart recording. An example will illustrate this concept.

Example 2.5 Consider a somewhat trivial process defined to be a constant with time, the constant being a random variable with zero-mean normal distri-

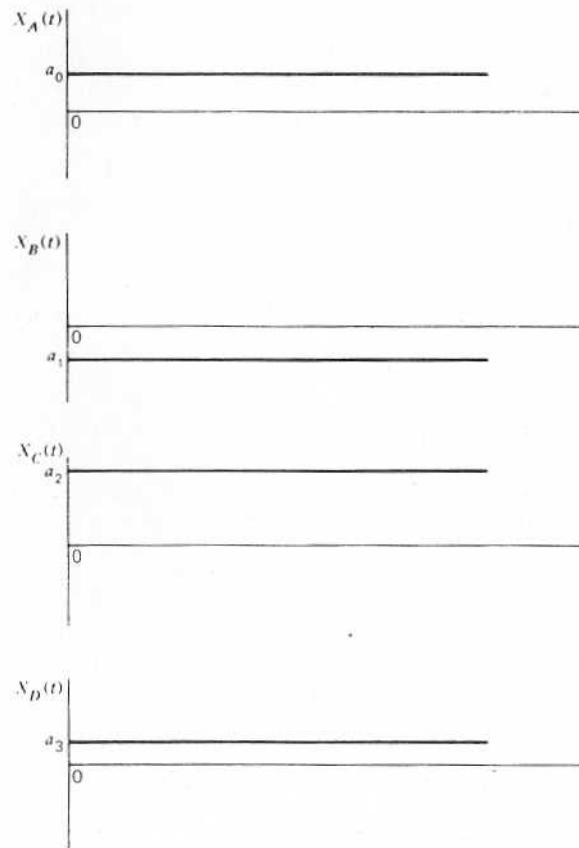


Figure 2.6. Ensemble of random constants.

bution. An ensemble of sample realizations for this process is shown in Fig. 2.6. A common physical situation where this kind of process model would be appropriate is random instrument bias. In many applications, some small residual random bias will remain in spite of all attempts to eliminate it, and the bias will be different for each instrument in the batch. In Fig. 2.6 we see that time samples collected from a single sample signal, say the first one, will all have the same value a_0 . The average of these is, of course, just a_0 . On the other hand, if we were to collect samples in an ensemble sense, the values $a_0, a_1, a_2, \dots, a_n$, would be obtained. These would have a normal distribution with zero mean. Obviously time and ensemble sampling do not lead to the same result in this case, so the process is not ergodic. It is, however, a stationary process because the "statistics" of the process do not change with time. ■

In the case of physical noise processes, one can rarely justify strict stationarity or ergodicity in a formal sense. Thus, we must often lean on heuristic knowledge of the processes involved and simply make assumptions accordingly.

Random processes are sometimes classified according to two categories, *deterministic* and *nondeterministic*. As might be expected, a deterministic random process resembles a deterministic nonrandom signal in that it has some special deterministic structure. Specifically, if the process description is such that knowledge of a sample signal's past enables exact prediction of its future, it is classified as a deterministic random process. Examples are:

- (1) $X(t) = a$; a is normal, $N(m, \sigma^2)$.
- (2) $X(t) = A \sin \omega t$; A is Rayleigh distributed.
- (3) $X(t) = A \sin(\omega t + \theta)$; A and θ are independent and Rayleigh and uniformly distributed, respectively.

In each case, if one were to specify a particular sample signal prior to some time, say t_1 , the sample realizations for that particular signal would be indirectly specified, and the signal's future values would be exactly predictable.

Random processes that are not deterministic are classified as nondeterministic. These processes have no special functional structure that enables their prediction by specification of certain key parameters or their past history. Typical "noise" is a good example of a nondeterministic random process. It wanders on aimlessly, as determined by chance, and has no particular deterministic structure.

2.5 AUTOCORRELATION FUNCTION

The autocorrelation function for a random process $X(t)$ is defined as

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)] \quad (2.5.1)$$

where t_1 and t_2 are arbitrary sampling times. Clearly, it tells how well the process is correlated with itself at two different times. If the process is stationary, its probability density functions are invariant with time, and the autocorrelation function depends only on the time difference $t_2 - t_1$. Thus R_X reduces to a function of just the time difference variable τ , that is,

$$R_X(\tau) = E[X(t)X(t + \tau)] \quad (\text{Stationary case}) \quad (2.5.2)$$

where t_1 is now denoted as just t and t_2 is $(t + \tau)$. Stationarity assures us that the expectation is not dependent on t .

Note that the autocorrelation function is the ensemble average (i.e., expectation) of the product of $X(t_1)$ and $X(t_2)$; therefore, it can formally be written as

$$R_X(t_1, t_2) = E[X_1 X_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \quad (2.5.3)$$

where we are using the shortened notation $X_1 = X(t_1)$ and $X_2 = X(t_2)$. However, Eq. (2.5.3) is often not the simplest way of determining R_X because the joint density function $f_{X_1 X_2}(x_1, x_2)$ must be known explicitly in order to evaluate the integral. If the ergodic hypothesis applies, it is often easier to compute R_X as a time average rather than an ensemble average. An example will illustrate this.

Example 2.6 Consider the same process defined in Example 2.3. A typical sample signal for this process is shown in Fig. 2.7 along with the same signal shifted in time an amount τ . Now, the process under consideration in this case is ergodic, so we should be able to interchange time and ensemble averages. Thus, the autocorrelation function can be written as

$$\begin{aligned} R_X(\tau) &= \text{time average of } X_A(t) \cdot X_A(t + \tau) \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T X_A(t) X_A(t + \tau) dt \end{aligned} \quad (2.5.4)$$

It is obvious that when $\tau = 0$, the integral of Eq. (2.5.4) is just the mean square value of $X_A(t)$, which is $\frac{1}{3}$ in this case. On the other hand, when τ is unity or larger, there is no overlap of the correlated portions of $X_A(t)$ and $X_A(t + \tau)$, and

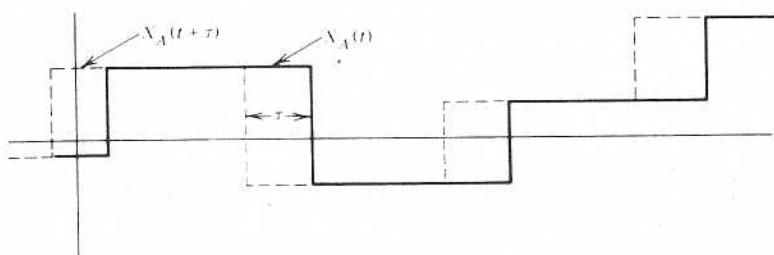


Figure 2.7. Random waveform for Example 2.6.

thus the average of the product is zero. Now, as the shift τ is reduced from 1 to 0, the overlap of correlated portions increases linearly until the maximum overlap occurs at $\tau = 0$. This then leads to the autocorrelation function shown in Fig. 2.8. Note that for stationary ergodic processes, the direction of time shift τ is immaterial, and hence the autocorrelation function is symmetric about the origin. Also, note that we arrived at $R_X(\tau)$ without formally finding the joint density function $f_{X_1 X_2}(x_1, x_2)$. ■

Sometimes, the random process under consideration is not ergodic, and it is necessary to distinguish between the usual autocorrelation function (ensemble average) and the time-average version. Thus, we define the *time autocorrelation function* as

$$\mathcal{R}_{X_A}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T X_A(t) X_A(t + \tau) dt \quad (2.5.4)$$

where $X_A(t)$ denotes a sample realization of the $X(t)$ process. There is the tacit assumption that the limit indicated in Eq. (2.5.4) exists. Also note that script \mathcal{R} rather than italic R is used as a reminder that this is a time average rather than an ensemble average.

Example 2.7 To illustrate the difference between the usual autocorrelation function and the time autocorrelation function, consider the deterministic random process

$$X(t) = A \sin \omega t \quad (2.5.5)$$

where A is a normal random variable with zero mean and variance σ^2 , and ω is a known constant. Suppose we obtain a single sample of A and its numerical value is A_1 . The corresponding sample of $X(t)$ would then be

$$X_A(t) = A_1 \sin \omega t \quad (2.5.6)$$

According to Eq. (2.5.4), its time autocorrelation function would then be

$$\begin{aligned} \mathcal{R}_{X_A}(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T A_1 \sin \omega t \cdot A_1 \sin \omega(t + \tau) dt \\ &= \frac{A_1^2}{2} \cos \omega \tau \end{aligned} \quad (2.5.7)$$

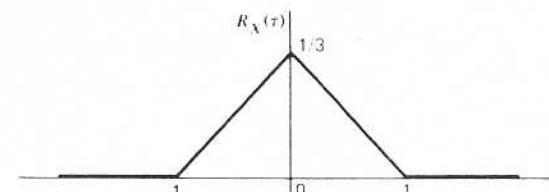


Figure 2.8. Autocorrelation function for Example 2.6.

On the other hand, the usual autocorrelation function is calculated as an ensemble average, that is, from Eq. (2.5.1). In this case it is

$$\begin{aligned} R_X(t_1, t_2) &= E[X(t_1)X(t_2)] \\ &= E[A \sin \omega t_1 \cdot A \sin \omega t_2] \\ &= \sigma^2 \sin \omega t_1 \sin \omega t_2 \end{aligned} \quad (2.5.8)$$

Note that this expression is quite different from that obtained for $R_{X_A}(\tau)$. Clearly, time averaging does not yield the same result as ensemble averaging, so the process is not ergodic. Furthermore, the autocorrelation function given by Eq. (2.5.8) does not reduce to simply a function of $t_2 - t_1$. Therefore, the process is not stationary. ■

General Properties of Autocorrelation Functions

There are some general properties that are common to all autocorrelation functions for stationary processes. These will now be enumerated with a brief comment about each:

1. $R_X(0)$ is the mean-square value of the process $X(t)$. This is self-evident from Eq. (2.5.2).
2. $R_X(\tau)$ is an even function of τ . This results from the stationarity assumption. [In the nonstationary case there is symmetry with respect to the two arguments t_1 and t_2 . In Eq. (2.5.1) it certainly makes no difference in which order we multiply $X(t_1)$ and $X(t_2)$. Thus, $R_X(t_1, t_2) = R_X(t_2, t_1)$.]
3. $|R_X(\tau)| \leq R_X(0)$ for all τ . We have assumed $X(t)$ is stationary and thus the means and mean-square values of $X(t)$ and $X(t + \tau)$ must be the same. Also the magnitude of the correlation coefficient relating two random variables is never greater than unity. Thus $R_X(\tau)$ can never be greater in magnitude than $R_X(0)$.
4. If $X(t)$ contains a periodic component, $R_X(\tau)$ will also contain a periodic component with the same period. This can be verified by writing $X(t)$ as the sum of the nonperiodic and periodic components and then applying the definition given by Eq. (2.5.2). It is of interest to note that if the process is ergodic as well as stationary and if the periodic component is sinusoidal, then $R_X(\tau)$ will contain no information about the phase of the sinusoidal component. The harmonic component always appears in the autocorrelation function as a cosine function, irrespective of its phase.
5. If $X(t)$ does not contain any periodic components, $R_X(\tau)$ tends to zero as $\tau \rightarrow \infty$. This is just a mathematical way of saying that $X(t + \tau)$ becomes completely uncorrelated with $X(t)$ for large τ if there are no hidden periodicities in the process. Note that a constant is a special case of a periodic function. Thus $R_X(\infty) = 0$ implies zero mean for the process.
6. The Fourier transform of $R_X(\tau)$ is real, symmetric, and non-negative. The

real, symmetric property follows directly from the even property of $R_X(\tau)$. The non-negative property is not obvious at this point. It will be justified later in Section 2.7, which deals with the spectral density function for the process.

It was mentioned previously that strict stationarity is a severe requirement, because it requires that all the higher-order probability density functions be invariant under a time translation. This is often difficult to verify. Thus, a less demanding form of stationarity is often used, or assumed. A random process is said to be *covariance stationary* or *wide-sense stationary* if $E[X(t_1)]$ is independent of t_1 and $E[X(t_1)X(t_2)]$ is dependent only on the time difference $t_2 - t_1$. Obviously, if the second-order density $f_{X_1X_2}(x_1, x_2)$ is independent of the time origin, the process is covariance stationary.

Further examples of autocorrelation functions will be given as this chapter progresses. We will see that the autocorrelation function is an important descriptor of a random process and one that is relatively easy to obtain because it only depends on the second-order probability density for the process.

2.6 CROSSCORRELATION FUNCTION

The crosscorrelation function between the processes $X(t)$ and $Y(t)$ is defined as

$$R_{XY}(t_1, t_2) = E[X(t_1)Y(t_2)] \quad (2.6.1)$$

Again, if the processes are stationary, only the time *difference* between sample points is relevant, so the crosscorrelation function reduces to

$$R_{XY}(\tau) = E[X(t)Y(t + \tau)] \quad (\text{Stationary case}) \quad (2.6.2)$$

Just as the autocorrelation function tells us something about how a process is correlated with itself, the crosscorrelation function provides information about the mutual correlation between two processes.

Notice that it is important to order the subscripts properly in writing $R_{XY}(\tau)$. A skew-symmetric relation exists for stationary processes as follows. By definition,

$$R_{XY}(\tau) = E[X(t)Y(t + \tau)] \quad (2.6.3)$$

$$R_{YX}(\tau) = E[Y(t)X(t + \tau)] \quad (2.6.4)$$

The expectation in Eq. (2.6.4) is invariant under a translation of $-\tau$. Thus, $R_{YX}(\tau)$ is also given by

$$R_{YX}(\tau) = E[Y(t - \tau)X(t)] \quad (2.6.5)$$

Now, comparing Eqs. (2.6.3) and (2.6.5), we see that

$$R_{XY}(\tau) = R_{YX}(-\tau) \quad (2.6.6)$$

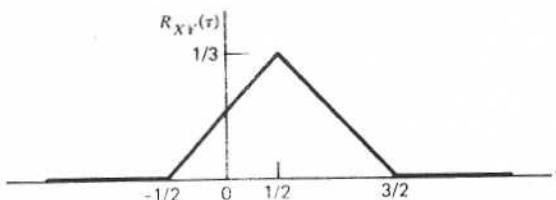


Figure 2.9. Crosscorrelation function for Example 2.8.

Thus, interchanging the order of the subscripts of the crosscorrelation function has the effect of changing the sign of the argument.

Example 2.8 Let $X(t)$ be the same random process of Example 2.6 and illustrated in Fig. 2.7. Let $Y(t)$ be the same signal as $X(t)$, but delayed one-half unit of time. The crosscorrelation $R_{XY}(\tau)$ would then be as shown in Fig. 2.9. Note that $R_{XY}(\tau)$ is not an even function of τ ; nor does its maximum occur at $\tau = 0$. Thus, the cross correlation function lacks the symmetry possessed by the autocorrelation function. ■

We frequently need to consider additive combinations of random processes. For example, let the process $Z(t)$ be the sum of stationary zero-mean processes $X(t)$ and $Y(t)$:

$$Z(t) = X(t) + Y(t) \quad (2.6.7)$$

The autocorrelation function of the summed process is then

$$\begin{aligned} R_Z(\tau) &= E\{[X(t) + Y(t)][X(t + \tau) + Y(t + \tau)]\} \\ &= E[X(t)X(t + \tau)] + E[Y(t)X(t + \tau)] \\ &\quad + E[X(t)Y(t + \tau)] + E[Y(t)Y(t + \tau)] \\ &= R_X(\tau) + R_{YX}(\tau) + R_{XY}(\tau) + R_Y(\tau) \end{aligned} \quad (2.6.8)$$

Now, if X and Y are uncorrelated processes, the middle terms of Eq. (2.6.8) are zero, and we have

$$R_Z(\tau) = R_X(\tau) + R_Y(\tau), \quad (\text{for } X \text{ and } Y \text{ uncorrelated}) \quad (2.6.9)$$

This can obviously be extended to the sum of more than two processes. Equation (2.6.9) is a much-used relationship, and it should always be remembered that it only applies when the processes being summed have zero crosscorrelation.

2.7 POWER SPECTRAL DENSITY FUNCTION

It was mentioned in Section 2.6 that the autocorrelation function is an important descriptor of a random process. Qualitatively, if the autocorrelation func-

tion decreases rapidly with τ , the process changes rapidly with time; conversely, a slowly changing process will have an autocorrelation function that decreases slowly with τ . Thus, we would suspect that this important descriptor contains information about the frequency content of the process; and this is in fact the case. For stationary processes, there is an important relation known as the *Wiener-Khinchine relation*:

$$S_X(j\omega) = \mathcal{F}[R_X(\tau)] = \int_{-\infty}^{\infty} R_X(\tau)e^{-j\omega\tau} d\tau \quad (2.7.1)$$

where $\mathcal{F}[\cdot]$ indicates Fourier transform and ω has the usual meaning of (2π) (frequency in hertz). S_X is called the *power spectral density function* or simply the *spectral density function* of the process.

The adjectives power and spectral come from the relationship of $S_X(j\omega)$ to the usual spectrum concept for a deterministic signal. However, some care is required in making this connection. If the process $X(t)$ is time stationary, it wanders on ad infinitum and is not absolutely integrable. Thus the defining integral for the Fourier transform does not converge. When considering the Fourier transform of the process, we are forced to consider a truncated version of it, say $X_T(t)$, which is truncated to zero outside a span of time T . The Fourier transform of a sample realization of the truncated process will then exist.

Let $\mathcal{F}\{X_T\}$ denote the Fourier transform of $X_T(t)$, where it is understood that for any given ensemble of samples of $X_T(t)$ there will be corresponding ensemble of $\mathcal{F}\{X_T(t)\}$. That is, $\mathcal{F}\{X_T(t)\}$ has stochastic attributes just as does $X_T(t)$. Now look at the following expectation:

$$E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right]$$

For any particular sample realization of $X_T(t)$, the quantity inside the brackets is known as the *periodogram* for that particular signal. It will now be shown that averaging over an ensemble of periodograms yields the power spectral density function.

The expectation of the periodogram of a signal spanning the time interval $[0, T]$ can be manipulated as follows:

$$\begin{aligned} &E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right] \\ &= E\left[\frac{1}{T} \int_0^T X(t)e^{-j\omega t} dt \int_0^T X(s)e^{j\omega s} ds\right] \\ &= \frac{1}{T} \int_0^T \int_0^T E[X(t)X(s)]e^{-j\omega(t-s)} dt ds \end{aligned} \quad (2.7.2)$$

Note that we were able to drop the subscript T on $X(t)$ because of the restricted range of integration. If we now assume $X(t)$ is stationary, $E[X(t)X(s)]$ becomes $R_X(t-s)$ and Eq. (2.7.2) becomes

$$E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right] = \frac{1}{T} \int_0^T \int_0^T R_X(t-s) e^{-j\omega(t-s)} ds dt \quad (2.7.3)$$

The appearance of $t - s$ in two places in Eq. (2.7.3) suggests a change of variables. Let

$$\tau = t - s \quad (2.7.4)$$

Equation (2.7.3) then becomes

$$\frac{1}{T} \int_0^T \int_0^T R_X(t-s) e^{-j\omega(t-s)} ds dt = -\frac{1}{T} \int_0^T \int_t^{t-T} R_X(\tau) e^{-j\omega\tau} d\tau dt \quad (2.7.5)$$

The new region of integration in the τt plane is shown in Fig. 2.10.

Next we interchange the order of integration and integrate over the two triangular regions separately. This leads to

$$\begin{aligned} E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right] &= \frac{1}{T} \int_{-T}^0 \int_0^{t+T} R_X(\tau) e^{-j\omega\tau} dt d\tau + \frac{1}{T} \int_0^T \int_\tau^T R_X(\tau) e^{-j\omega\tau} dt d\tau \end{aligned} \quad (2.7.6)$$

We now integrate with respect to t with the result

$$\begin{aligned} E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right] &= \frac{1}{T} \int_{-T}^0 (\tau + T) R_X(\tau) e^{-j\omega\tau} d\tau + \frac{1}{T} \int_0^T (T - \tau) R_X(\tau) e^{-j\omega\tau} d\tau \end{aligned} \quad (2.7.7)$$

Finally, Eq. (2.7.7) may be written in more compact form as

$$E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right] = \int_{-T}^T \left(1 - \frac{|\tau|}{T}\right) R_X(\tau) e^{-j\omega\tau} d\tau \quad (2.7.8)$$

The factor $1 - |\tau|/T$ that multiplies $R_X(\tau)$ may be thought of as a triangular weighting factor that approaches unity as T becomes larger; at least this is true if $R_X(\tau)$ approaches zero as τ becomes larger, which it will do if $X(t)$ contains no periodic components. Thus, as T becomes large, we have the following relationship:

$$E\left[\frac{1}{T}|\mathcal{F}\{X_T(t)\}|^2\right] \Rightarrow \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau, \quad \text{as } T \rightarrow \infty \quad (2.7.9)$$

Or, in other words,

$$\text{Average periodogram} \Rightarrow \text{power spectral density function} \quad (2.7.10)$$

as T becomes larger. Equation (2.7.9) is a most important relationship, because it is this that ties the spectral function $S_X(j\omega)$ to "spectrum" as thought of in the usual deterministic sense. Remember that the spectral density function, as formally defined by Eq. (2.7.1), is a probabilistic concept. On the other hand, the periodogram is a spectral concept in the usual sense of being related to the

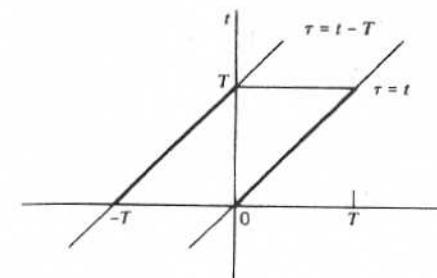


Figure 2.10. Region of integration in the τt plane.

Fourier transform of a time signal. The relationship given by Eq. (2.7.9) then provides the tie between the probabilistic and spectral descriptions of the process, and it is this equation that suggests the name for $S_X(j\omega)$, power spectral density function. More will be said of this in Section 2.14, which deals with the determination of the spectral function from experimental data.

Because of the special attributes of the autocorrelation function $R_X(\tau)$, its Fourier transform $S_X(j\omega)$ always works out to be a real, nonnegative, symmetric function of ω . This should be apparent from the left side of Eq. (2.7.9), and will be illustrated in Example 2.9.

Example 2.9 Consider a random process $X(t)$ whose autocorrelation function is given by

$$R_X(\tau) = \sigma^2 e^{-\beta|\tau|} \quad (2.7.11)$$

where σ^2 and β are known constants. The spectral density function for the $X(t)$ process is

$$S_X(j\omega) = \mathcal{F}[R_X(\tau)] = \frac{\sigma^2}{j\omega + \beta} + \frac{\sigma^2}{-j\omega + \beta} = \frac{2\sigma^2\beta}{\omega^2 + \beta^2} \quad (2.7.12)$$

Both R_X and S_X are sketched in Fig. 2.11. ■

Occasionally it is convenient to write the spectral density function in terms of the complex frequency variable s rather than ω . This is done by simply

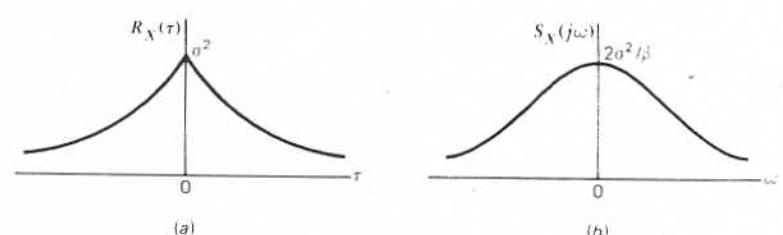


Figure 2.11. Autocorrelation and spectral density functions for Example 2.9. (a) Autocorrelation function. (b) Spectral function.

replacing $j\omega$ with s ; or, equivalently, replacing ω^2 with $-s^2$. For Example 2.9, the spectral density function in terms of s is then

$$S_X(s) = \frac{2\sigma^2\beta}{\omega^2 + \beta^2} \Big|_{\omega^2 = -s^2} = \frac{2\sigma^2\beta}{-s^2 + \beta^2} \quad (2.7.13)$$

It should be clear now why we chose to include the "j" with ω in $S_X(j\omega)$, even though $S_X(j\omega)$ always works out to be a real function of ω . By writing the argument of $S_X(j\omega)$ as $j\omega$, rather than just ω , we can use the same symbol for spectral function in either the complex or real frequency domain. That is,

$$S_X(s) = S_X(j\omega) \quad (2.7.14)$$

is correct notation in the usual mathematical sense.

From Fourier transform theory, we know that the inverse transform of the spectral function should yield the autocorrelation function, that is,

$$\mathcal{F}^{-1}[S_X(j\omega)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(j\omega) e^{j\omega\tau} d\omega = R_X(\tau) \quad (2.7.15)$$

If we let $\tau = 0$ in Eq. (2.7.15), we get

$$R_X(0) = E[X^2(t)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(j\omega) d\omega \quad (2.7.16)$$

Equation (2.7.16) provides a convenient means of computing the mean square value of a stationary process, given its spectral function. As mentioned before, it is sometimes convenient to use the complex frequency variable s rather than $j\omega$. If this is done, Eq. (2.7.16) becomes

$$E[X^2(t)] = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} S_X(s) ds \quad (2.7.17)$$

Equation (2.7.16) suggests that we can consider the signal power as being distributed in frequency in accordance with $S_X(j\omega)$; thus the terms power and density in power spectral density function. Using this concept, we can obtain the power in a finite band by integrating over the appropriate range of frequencies, that is,

$$\left[\text{"Power" in range } \omega_1 \leq \omega \leq \omega_2 \right] = \frac{1}{2\pi} \int_{-\omega_2}^{-\omega_1} S_X(j\omega) d\omega + \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} S_X(j\omega) d\omega \quad (2.7.18)$$

An example will now be given to illustrate the use of Eqs. (2.7.16) and (2.7.17).

Example 2.10 Consider the spectral function of Example 2.9:

$$S_X(j\omega) = \frac{2\sigma^2\beta}{\omega^2 + \beta^2} \quad (2.7.19)$$

Application of Eq. (2.7.16) should yield the mean square value σ^2 . This can be verified using conventional integral tables.

$$E(X^2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2\sigma^2\beta}{\omega^2 + \beta^2} d\omega = \frac{\sigma^2\beta}{\pi} \left[\frac{1}{\beta} \tan^{-1} \frac{\omega}{\beta} \right]_{-\infty}^{\infty} = \sigma^2 \quad (2.7.20)$$

Or, equivalently, in terms of s :

$$E(X^2) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{2\sigma^2\beta}{-s^2 + \beta^2} ds = \sigma^2 \quad (2.7.21)$$

More will be said about evaluating integrals of the type in Eq. (2.7.21) later in Chapter 3. ■

In summary, we see that the autocorrelation function and spectral density function are Fourier transform pairs. Thus, both contain the same basic information about the process, but in different forms. Since we can easily transform back and forth between the time and frequency domains, the manner in which the information is presented is purely a matter of convenience for the problem at hand.

2.8 CROSS SPECTRAL DENSITY FUNCTION

Cross Spectral Density functions for stationary processes $X(t)$ and $Y(t)$ are defined as

$$S_{XY}(j\omega) = \mathcal{F}[R_{XY}(\tau)] = \int_{-\infty}^{\infty} R_{XY}(\tau) e^{-j\omega\tau} d\tau \quad (2.8.1)$$

$$S_{YX}(j\omega) = \mathcal{F}[R_{YX}(\tau)] = \int_{-\infty}^{\infty} R_{YX}(\tau) e^{-j\omega\tau} d\tau \quad (2.8.2)$$

The crosscorrelation functions $R_{XY}(\tau)$ and $R_{YX}(\tau)$ are not necessarily even functions of τ , and thus the corresponding cross spectral densities are usually not real functions of ω . It was noted in Section 2.6 that $R_{XY}(\tau) = R_{YX}(-\tau)$. Thus S_{XY} and S_{YX} are complex conjugates of each other:

$$S_{XY}(j\omega) = S_{YX}^*(j\omega) \quad (2.8.3)$$

and the sum of S_{XY} and S_{YX} is real.

Another function that is closely related to the cross spectral density is the *coherence function*. It is defined as

$$\gamma_{XY}^2 = \frac{|S_{XY}(j\omega)|^2}{S_X(j\omega)S_Y(j\omega)} \quad (2.8.4)$$

The coherence function can be seen to be normalized, and it is sort of a "correlation coefficient in the frequency domain." To see the normalization, let

$X(t) = Y(t)$ (maximum correlation) and then

$$\gamma_{XY}^2 = \frac{|S_{XY}(j\omega)|^2}{S_X(j\omega)S_Y(j\omega)} = 1 \quad (2.8.5)$$

On the other extreme, if $X(t)$ and $Y(t)$ are uncorrelated, $S_{XY}(j\omega) = 0$ and $\gamma_{XY}^2 = 0$. The coherence function is useful in analysis of random data and the reader can refer to Bendat and Piersol (11) for details. (Also see Problem 3.24).

If $Z(t)$ is the sum of zero-mean processes $X(t)$ and $Y(t)$, the spectral density of $Z(t)$ is given by

$$S_Z(j\omega) = \mathcal{F}[R_{X+Y}(\tau)] \quad (2.8.6)$$

Referring to Eq. (2.6.8), we then have

$$S_Z(j\omega) = S_X(j\omega) + S_{XY}(j\omega) + S_{YX}(j\omega) + S_Y(j\omega) \quad (2.8.7)$$

Just as in the case of the autocorrelation function, the two middle terms in Eq. (2.8.7) are zero if the X and Y processes are uncorrelated. So, for this special case,

$$S_{X+Y}(j\omega) = S_X(j\omega) + S_Y(j\omega), \quad (\text{for } X \text{ and } Y \text{ uncorrelated}) \quad (2.8.8)$$

2.9 WHITE NOISE

White noise is defined to be a stationary random process having a constant spectral density function. The term "white" is an obvious carryover from optics where white light is light containing all visible frequencies. Denoting the white-noise spectral amplitude as A , we then have

$$S_{wn}(j\omega) = A \quad (2.9.1)$$

The corresponding autocorrelation function for white noise is then

$$R_{wn}(\tau) = A\delta(\tau) \quad (2.9.2)$$

In analysis, one frequently makes simplifying assumptions in order to make the problem mathematically tractable. White noise is a good example of this. However, by assuming the spectral amplitude of white noise to be constant for all frequencies (for the sake of mathematical simplicity), we find ourselves in the awkward situation of having defined a process with infinite variance. Qualitatively, white noise is sometimes characterized as noise that is jumping around infinitely far, infinitely fast! This is obviously physical nonsense but it is a useful abstraction. The saving feature is that all physical systems are bandlimited to some extent, and a bandlimited system driven by white noise yields a process that has finite variance; that is, the end result makes sense. We will elaborate on this further in Chapter 3.

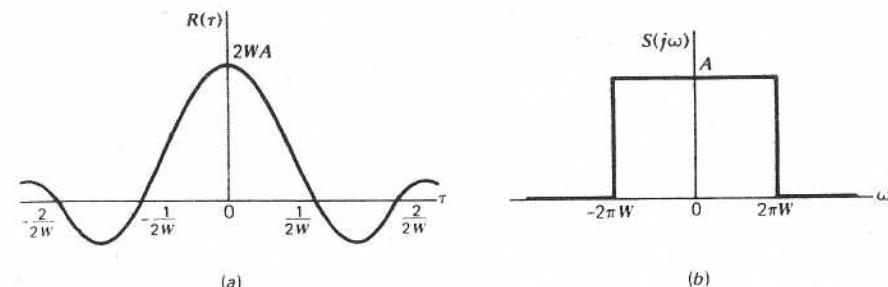


Figure 2.12. Baseband bandlimited white noise. (a) Autocorrelation function. (b) Spectral density function.

Bandlimited white noise is a random process whose spectral amplitude is constant over a finite range of frequencies, and zero outside that range. If the bandwidth includes the origin (sometimes called baseband), we then have

$$S_{bwn}(j\omega) = \begin{cases} A, & |\omega| \leq 2\pi W \\ 0, & |\omega| > 2\pi W \end{cases} \quad (2.9.3)$$

where W is the physical bandwidth in hertz. The corresponding autocorrelation function is

$$R_{bwn}(\tau) = 2WA \frac{\sin(2\pi W\tau)}{2\pi W\tau} \quad (2.9.4)$$

Both the autocorrelation and spectral density functions for baseband bandlimited white noise are sketched in Fig. 2.12. It is of interest to note that the autocorrelation function for baseband bandlimited white noise is zero for $\tau = 1/2W, 2/2W, 3/2W$, etc. From this we see that if the process is sampled at a rate of $2W$ samples/second (sometimes called the Nyquist rate), the resulting set of random variables are uncorrelated. Since this usually simplifies the analysis, the white bandlimited assumption is frequently made in bandlimited situations.

The frequency band for bandlimited white noise is sometimes offset from the origin and centered about some center frequency W_0 . It is easily verified that the autocorrelation/spectral-function pair for this situation is

$$S(j\omega) = \begin{cases} A, & 2\pi W_1 \leq |\omega| \leq 2\pi W_2 \\ 0, & |\omega| < 2\pi W_1 \text{ and } |\omega| > 2\pi W_2 \end{cases} \quad (2.9.5)$$

$$\begin{aligned} R(\tau) &= A \left[2W_2 \frac{\sin 2\pi W_2 \tau}{2\pi W_2 \tau} - 2W_1 \frac{\sin 2\pi W_1 \tau}{2\pi W_1 \tau} \right] \\ &= A2 \Delta W \frac{\sin \pi \Delta W \tau}{\pi \Delta W \tau} \cos 2\pi W_0 \tau \end{aligned} \quad (2.9.6)$$

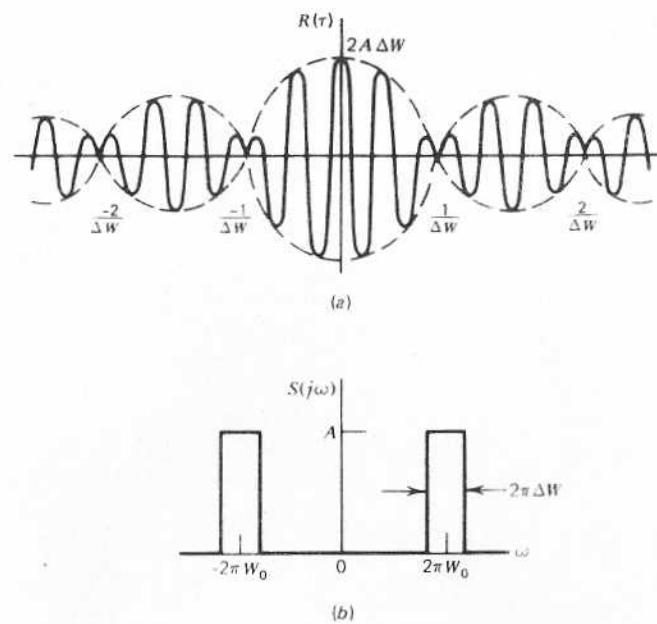


Figure 2.13. Bandlimited white noise with center frequency W_0 . (a) Autocorrelation function. (b) Spectral density.

where

$$\Delta W = W_2 - W_1 \text{ Hz}$$

$$W_0 = \frac{W_1 + W_2}{2} \text{ Hz}$$

These functions are sketched in Fig. 2.13.

It is worth noting the bandlimited white noise has a finite mean-square value, and thus it is physically plausible, whereas pure white noise is not. However, the mathematical forms for the autocorrelation and spectral functions in the bandlimited case are more complicated than for pure white noise.

2.10 GAUSS-MARKOV PROCESS

A stationary Gaussian process $X(t)$ that has an exponential autocorrelation is called a *Gauss-Markov* process. The autocorrelation and spectral functions for this process are then of the form

$$R_X(\tau) = \sigma^2 e^{-\beta|\tau|} \quad (2.10.1)$$

$$S_X(j\omega) = \frac{2\sigma^2\beta}{\omega^2 + \beta^2} \quad \left[\text{or } S_X(s) = \frac{2\sigma^2\beta}{-s^2 + \beta^2} \right] \quad (2.10.2)$$

These functions are sketched in Fig. 2.14. The mean-square value and time constant for the process are given by the σ^2 and $1/\beta$ parameters, respectively. The process is nondeterministic, so a typical sample time function would show no deterministic structure and would look like typical "noise." The exponential autocorrelation function indicates that sample values of the process gradually become less and less correlated as the time separation between samples increases. The autocorrelation function approaches zero as $\tau \rightarrow \infty$, and thus the mean value of the process must be zero. The reference to Markov in the name of this process is not obvious at this point, but it will be after the discussion on optimal prediction in Chapter 4.

The Gauss-Markov process is an important process in applied work because (1) it seems to fit a large number of physical processes with reasonable accuracy, and (2) it has a relatively simple mathematical description. As in the case of all stationary Gaussian processes, *specification of the process autocorrelation function completely defines the process*. This means that any desired higher-order probability density function for the process may be written out explicitly, given the autocorrelation function. An example will illustrate this.

Example 2.11 Let us say that a Gauss-Markov process $X(t)$ has autocorrelation function

$$R_X(\tau) = 100e^{-2|\tau|} \quad (2.10.3)$$

We wish to write out the third-order probability density function

$$f_{X_1 X_2 X_3}(x_1, x_2, x_3) \quad \text{where } X_1 = X(0), X_2 = X(.5), \text{ and } X_3 = X(1)$$

First we note that the process mean is zero. The covariance matrix in this case is a 3×3 matrix and is obtained as follows:

$$\begin{aligned} \mathbf{C}_X &= \begin{bmatrix} E(X_1^2) & E(X_1 X_2) & E(X_1 X_3) \\ E(X_2 X_1) & E(X_2^2) & E(X_2 X_3) \\ E(X_3 X_1) & E(X_3 X_2) & E(X_3^2) \end{bmatrix} = \begin{bmatrix} R_X(0) & R_X(.5) & R_X(1) \\ R_X(.5) & R_X(0) & R_X(.5) \\ R_X(1) & R_X(.5) & R_X(0) \end{bmatrix} \\ &= \begin{bmatrix} 100 & 100e^{-1} & 100e^{-2} \\ 100e^{-1} & 100 & 100e^{-1} \\ 100e^{-2} & 100e^{-1} & 100 \end{bmatrix} \end{aligned} \quad (2.10.4)$$

Now that \mathbf{C}_X has been written out explicitly, we can use the general normal form given by Eq. (1.15.5). The desired density function is then

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{3/2} |\mathbf{C}_X|^{1/2}} e^{-\frac{1}{2} \mathbf{x}^T \mathbf{C}_X^{-1} \mathbf{x}} \quad (2.10.5)$$

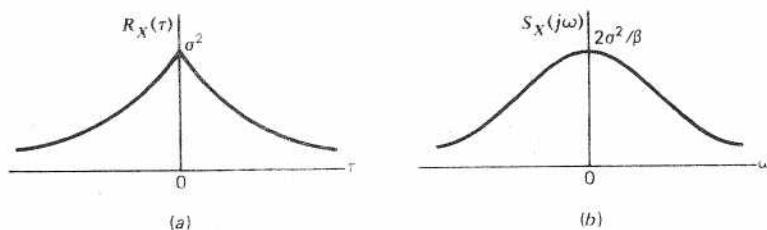


Figure 2.14. Autocorrelation and spectral density functions for Gauss-Markov process.
(a) Autocorrelation function. (b) Spectral density.

where \mathbf{x} is the 3-tuple

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (2.10.6)$$

and \mathbf{C}_X is given by Eq. (2.10.4). ■

2.11 RANDOM TELEGRAPH WAVE

Consider a binary voltage waveform that is generated according to the following rules:

- (a) The voltage is either +1 or -1 V.
- (b) The state at $t = 0$ may be either +1 or -1 V with equal likelihood.
- (c) As time progresses, the voltage switches from one state to the other at random. Specifically, the probability of k switches in a time interval T is governed by the Poisson distribution

$$P(k) = \frac{(aT)^k}{k!} e^{-aT} \quad (2.11.1)$$

where a is the average number of switches per unit time.

This random process is called the random telegraph wave, and a sample waveform is shown in Fig. 2.15. It is worth noting that the likelihood of switching at any point in time does not depend on the particular state or the length of time the system has been in that state.

A rigorous derivation of the autocorrelation function for the random telegraph wave can be found in a number of references (5,6). For our purposes here, we can use the following heuristic argument. Consider the product of successive time samples $X(t_1)$ and $X(t_2)$; the result must be either of two possibilities, +1 or -1. If the time interval $t_2 - t_1$ is small, $X(t_1)$ and $X(t_2)$ are highly

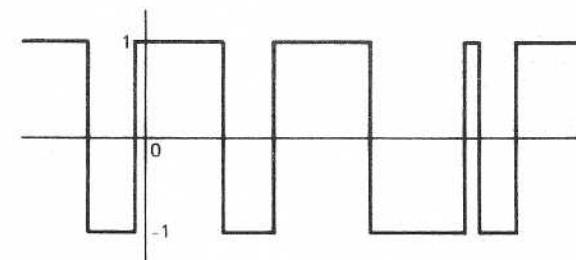


Figure 2.15. Random telegraph wave.

correlated, so that $X(t_1)X(t_2)$ is nearly unity. Then, as the spacing between samples is increased, their correlation gradually decreases and approaches zero as the spacing between samples goes to infinity. This leads to an exponential autocorrelation function of the form

$$R_X(\tau) = e^{-2a|\tau|} \quad (2.11.2)$$

where a is the average number of switches per unit time.

It should be apparent that the random telegraph wave is not a Gaussian process—far from it! Yet the Gauss-Markov process described in Section 2.10 has an autocorrelation function identical in form to that of the random telegraph wave. For purposes of comparison, a typical Gauss-Markov signal with about the same time constant as the random telegraph wave of Fig. 2.15 is shown in Fig. 2.16. The difference is striking. This is a vivid example of two processes that have radically different random structures, but still have the same autocorrelation functions and, of course, identical spectral characteristics. The moral is this:



Figure 2.16. Gauss-Markov signal with about the same time constant as for the random telegraph wave of Figure 2.15.

The autocorrelation function and/or spectral characteristics do not tell the whole story; all the probability density functions must be specified in order for the process to be completely described. In the Gauss-Markov case, this was done by specifying “Gaussian process” in addition to the autocorrelation function. In the random telegraph wave case, the higher-order densities were indirectly specified by describing a conceptual chance experiment creating the waveform. Obviously the probability density functions are radically different for the two processes in this case.

2.12 NARROWBAND GAUSSIAN PROCESS

In both control and communication systems, we frequently encounter situations where a very narrowband system is excited by wideband Gaussian noise. A high-Q tuned circuit or a lightly damped mass-spring arrangement are examples of narrowband systems. The resulting output is a noise process with essentially all its spectral content concentrated in a narrow frequency range. If one were to observe the output of such a system, the time function would appear to be nearly sinusoidal, especially if just a few cycles of the output signal were observed. However, if one were to carefully examine a long record of the signal, it would be seen that the quasi-sinusoid is slowly varying in both amplitude and phase. Such a signal is called narrowband noise and, if it is the result of passing wideband Gaussian noise through a linear narrowband system, then it is also Gaussian. We are assured of this because any linear operation on a set of normal variates results in another set of normal variates. The quasi-sinusoidal character depends only on the narrowband property, and the exact spectral shape within the band is immaterial.

The mathematical description of narrowband Gaussian noise follows. We first write the narrowband signal as

$$S(t) = X(t) \cos \omega_c t - Y(t) \sin \omega_c t \quad (2.12.1)$$

where $X(t)$ and $Y(t)$ are independent Gaussian random processes with similar narrowband spectral functions that are centered about zero frequency. The frequency ω_c is usually called the carrier frequency, and the effect of multiplying $X(t)$ and $Y(t)$ by $\cos \omega_c t$ and $\sin \omega_c t$ is to translate the baseband spectrum up to a similar spectrum centered about ω_c (see Problem 2.30). The independent $X(t)$ and $Y(t)$ processes are frequently called the in-phase and quadrature components of $S(t)$. Now, think of time t as a particular time, and think of $X(t)$ and $Y(t)$ as the corresponding random variables. Then make the usual rectangular to polar transformation via the equations

$$\begin{aligned} X &= R \cos \Theta \\ Y &= R \sin \Theta \end{aligned} \quad (2.12.2)$$

or, equivalently,

$$\begin{aligned} R &= \sqrt{X^2 + Y^2} \\ \Theta &= \tan^{-1} \frac{Y}{X} \end{aligned} \quad (2.12.3)$$

By substituting Eq. (2.12.2) into Eq. (2.12.1), we can now write $S(t)$ in the form

$$\begin{aligned} S(t) &= R(t) \cos \Theta(t) \cos \omega_c t - R(t) \sin \Theta(t) \sin \omega_c t \\ &= R(t) \cos [\omega_c t + \Theta(t)] \end{aligned} \quad (2.12.4)$$

It is from Eq. (2.12.4) that we get the physical interpretation of "slowly varying envelope (amplitude) and phase."

Before proceeding, a word or two about the probability densities for X , Y , R , and Θ is in order. If X and Y are independent normal random variables with the same variance σ^2 , their individual and joint densities are

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \quad (2.12.5)$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} \quad (2.12.6)$$

and

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2.12.7)$$

The corresponding densities for R and Θ are Rayleigh and uniform (see Example 1.24). The mathematical forms are

$$f_R(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \quad r \geq 0 \quad (\text{Rayleigh}) \quad (2.12.8)$$

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 \leq \theta < 2\pi \\ 0, & \text{otherwise} \end{cases} \quad (\text{Uniform}) \quad (2.12.9)$$

Also, the joint density function for R and Θ is

$$f_{R\Theta}(r, \theta) = \frac{r}{2\pi\sigma^2} e^{-r^2/2\sigma^2}, \quad r \geq 0 \quad \text{and} \quad 0 \leq \theta < 2\pi \quad (2.12.10)$$

It is of interest to note here that if we consider simultaneous time samples of envelope and phase, the resulting random variables are statistically independent. However, the processes $R(t)$ and $\Theta(t)$ are not statistically independent (5). This is due to the fact that the joint probability density associated with adjacent samples cannot be written in product form, that is,

$$f_{R_1 R_2 \Theta_1 \Theta_2}(r_1, r_2, \theta_1, \theta_2) \neq f_{R_1 R_2}(r_1, r_2) f_{\Theta_1 \Theta_2}(\theta_1, \theta_2) \quad (2.12.11)$$

We have assumed that $S(t)$ is a Gaussian process, and from Eq. (2.12.1) we see that

$$\text{Var } S = \frac{1}{2} (\text{Var } X) + \frac{1}{2} (\text{Var } Y) = \sigma^2 \quad (2.12.12)$$

Thus

$$f_S(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-s^2/2\sigma^2} \quad (2.12.13)$$

The higher-order density functions for S will, of course, depend on the specific shape of the spectral density for the process.

2.13 WIENER OR BROWNIAN-MOTION PROCESS

Suppose we start at the origin and take n steps forward or backward at random, with equal likelihood of stepping in either direction. We pose two questions: After taking n steps, (a) what is the average distance traveled, and (b) what is the variance of the distance? This is the classical random-walk problem of statistics. The averages considered here must be taken in an ensemble sense; for example, think of running simultaneous experiments and then averaging the results for a given number of steps. It should be apparent that the average distance traveled is zero, provided we say "forward" is positive and "backward" is negative. However, the square of the distance is always positive (or zero), so its average for a large number of trials will not be zero. It is shown in elementary statistics that the variance after n unit steps is just n , or the standard deviation is \sqrt{n} (see Problem 2.21). Note that this increases without bound as n increases, and thus the process is nonstationary.

The continuous analog of random-walk is the output of an integrator driven with white noise. This is shown in block-diagram form in Fig. 2.17a. Here we consider the input switch as closing at $t = 0$ and the initial integrator output as being zero. An ensemble of typical output time signals is shown in Fig. 2.17b. The system response $X(t)$ is given by

$$X(t) = \int_0^t F(u) du \quad (2.13.1)$$

Clearly, the average of the output is

$$E[X(t)] = E \left[\int_0^t F(u) du \right] = \int_0^t E[F(u)] du = 0 \quad (2.13.2)$$

Also, the mean-square-value (variance) is

$$E[X^2(t)] = E \left[\int_0^t F(u) du \int_0^t F(v) dv \right] = \int_0^t \int_0^t E[F(u)F(v)] du dv \quad (2.13.3)$$

But $E[F(u)F(v)]$ is just the autocorrelation function $R_F(u - v)$, which in this case is a Dirac delta function. Thus

$$E[X^2(t)] = \int_0^t \int_0^t \delta(u - v) du dv = \int_0^t dv = t \quad (2.13.4)$$

Thus, $E[X^2(t)]$ increases linearly with time and the rms value increases in accordance with \sqrt{t} (for unity white noise input).

Now, add the further requirement that the input be *Gaussian* white noise. The output will then be a Gaussian process because integration is a linear operation on the input. The resulting continuous random-walk process is known as the *Wiener* or *Brownian-motion* process. The process is nonsta-

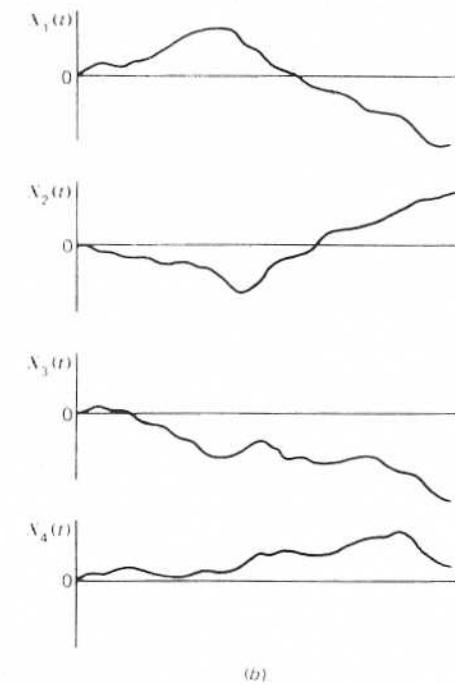
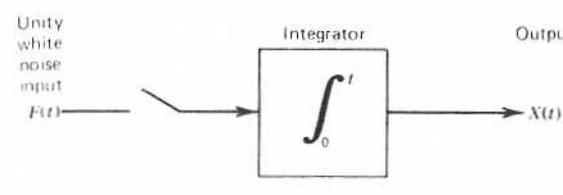


Figure 2.17. Continuous analog of random walk. (a) Block diagram. (b) Ensemble of output signals.

tionary, it is Gaussian, and its mean, mean-square value, and autocorrelation function are given by

$$E[X(t)] = 0 \quad (2.13.5)$$

$$E[X^2(t)] = t \quad (2.13.6)$$

$$\begin{aligned} R_X(t_1, t_2) &= E[X(t_1)X(t_2)] = E \left[\int_0^{t_1} F(u) du \cdot \int_0^{t_2} F(v) dv \right] \\ &= \int_0^{t_1} \int_0^{t_2} E[F(u)F(v)] du dv = \int_0^{t_1} \int_0^{t_2} \delta(u - v) du dv \end{aligned}$$

Evaluation of the double integral yields

$$R_X(t_1, t_2) = \begin{cases} t_2, & t_1 \geq t_2 \\ t_1, & t_1 \leq t_2 \end{cases} \quad (2.13.7)$$

Since the process is nonstationary, the autocorrelation function is a general function of the two arguments t_1 and t_2 . With a little imagination, Eq. (2.13.7) can be seen to describe two faces of a pyramid with the sloping ridge of the pyramid running along the line $t_1 = t_2$.

It was mentioned before that there are difficulties in defining directly what is meant by Gaussian white noise. This is because of the "infinite variance" problem. The Wiener process is well-behaved, though. Thus we can reverse the argument given here and begin by arbitrarily defining it as a Gaussian process with an autocorrelation function given by Eq. (2.13.7). This completely specifies the process. We can now describe Gaussian white noise in terms of its integral. That is, Gaussian white noise is that hypothetical process which, when integrated, yields a Wiener process.

2.14 PSEUDORANDOM SIGNALS

As the name implies, pseudorandom signals have the appearance of being random, but are not truly random. In order for a signal to be truly random, there must be some uncertainty about it that is governed by chance. Pseudorandom signals do not have this "chance" property. Two examples of pseudorandom signals will now be presented.

Example 2.12 Consider a sample realization of finite length T of a Gaussian-Markov process. Let the time length T of the sample be large relative to the time constant of the process. After the sample is taken, of course, the time function in all its intimate detail is known to the observer. After the fact, nothing remains to chance insofar as the observer is concerned. Next, imagine folding this record back on itself into a single loop (it might be on magnetic tape), and then imagine playing the loop continuously. It should be clear that the resulting signal would be periodic and completely known (determined), at least to the original observer. Yet, to a second observer casually looking at a small portion of the loop, the signal would appear to be just random noise. It should be mentioned that this is not a completely hypothetical situation: experimental spectral analysis was frequently implemented in just this way in times prior to modern on-line digital methods.

The "looped" signal that goes on ad infinitum is periodic, so it would have line-type rather than continuous spectral characteristics. See Fig. 2.18. ■

Line-type spectra are characteristic of all pseudorandom signals. The line spacing may be extremely small, as is the case for very large T , but it is there.

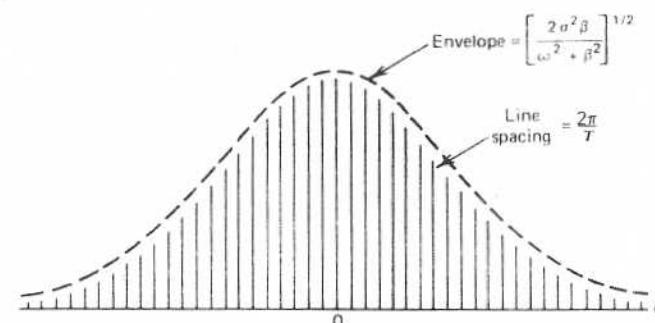


Figure 2.18. Spectrum of noise sample looped back on itself

nevertheless. Note that the envelope of the lines would approximate the square root of the spectral density of the process from which the sample was taken, provided the record length T is large.* In this case the usual laboratory analog spectrum analyzer would not be able to resolve individual lines, and it would indicate a smooth spectrum proportional to the line envelope. This would be just fine and what was desired if the original experimental problem was to determine the spectral characteristics of random process from a single long sample of the process. The point of all this is that the typical analog spectrum analyzer could not distinguish between pseudorandom noise and true random noise if the line spacing for the pseudorandom noise is very small.

Example 2.13 Binary sequences generated by shift registers with feedback have periodic properties and have found extensive application in ranging and communication systems (7,8,9). We will use the simple 5-bit shift register shown in Fig. 2.19 to demonstrate how a pseudorandom binary signal can be generated. In this system the bits are shifted to the right with each clock pulse, and the input on the left is determined by the feedback arrangement. For the initial condition shown, it can be verified that the output sequence is

11110001101110101000100101100111100

Note that the sequence repeats itself after 31 bits. This periodic property is characteristic of a shift register with feedback. The maximum length of the sequence (before repetition) is given by $(2^n - 1)$ where n is the register length (7). The 5-bit example used here is then a maximum-length sequence. These sequences are especially interesting because of their pseudorandom appearance. Note that there are nearly the same number of zeros and ones in the sequence (16 ones and 15 zeros), and that they appear to occur more or less at

* Strictly speaking, the *average* envelope would approximate the square root of the spectral density (see Section 2.15).

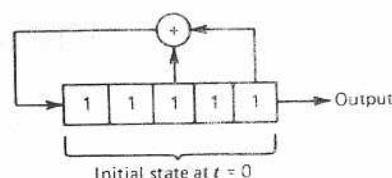


Figure 2.19. Binary shift register with feedback.

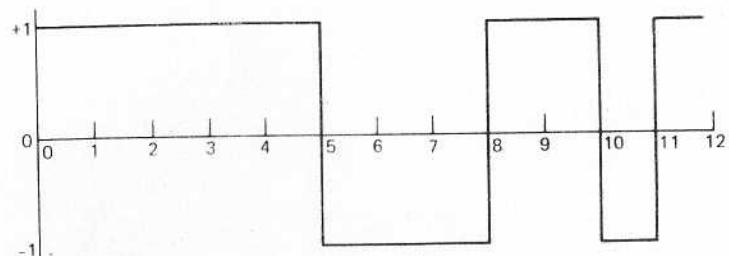


Figure 2.20. Pseudorandom binary waveform.

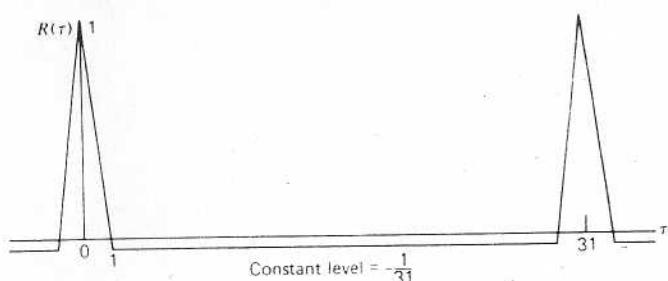


Figure 2.21. Time autocorrelation function for waveform of Figure 2.20.

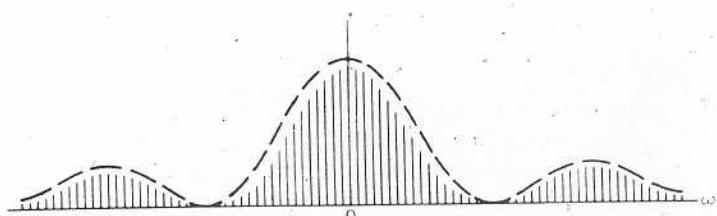


Figure 2.22. Spectral density for pseudorandom binary waveform.

random. If we consider a longer shift register, say one with 10 bits, its maximum-length sequence would be 1023 bits. It would have 512 ones and 511 zeros; and, again, these would appear to be distributed at random. A casual look at any interior string of bits would not reveal anything systematic. Yet the string of bits so generated is entirely deterministic. Once the feedback arrangement and initial condition are specified, the output is determined forever after. Thus, the sequence is pseudorandom, not random. ■

When converted to voltage waveforms, maximum-length sequences also have special autocorrelation properties. Returning to the 31-bit example, let binary one be 1 V and binary zero be -1 V, and let the voltage level be held constant during the clock-pulse interval. The resulting time waveform is shown in Fig. 2.20, and its time autocorrelation function is shown in Fig. 2.21. Note that the unique distribution of zeros and ones for this sequence is such that the autocorrelation function is a small constant value after a shift of one unit of time (i.e., one bit). This is typical of all maximum-length sequences. When the sequence length is long, the correlation after a shift of one unit is quite small. This has obvious advantages in correlation detection schemes, and such schemes have been used extensively in electronic ranging applications (7,8,9).

The spectral density function for the waveform of Fig. 2.20 is shown in Fig. 2.22. As with all pseudorandom signals, the spectrum is line-type rather than continuous.

2.15 DETERMINATION OF AUTOCORRELATION AND SPECTRAL DENSITY FUNCTIONS FROM EXPERIMENTAL DATA

The determination of spectral characteristics of a random process from experimental data is a common engineering problem. All of the optimization techniques presented in the following chapters depend on prior knowledge of the spectral density of the processes involved. Thus, the designer needs this information and it usually must come from experimental evidence. Spectral determination is a relatively complicated problem with many pitfalls, and one should approach it with a good deal of caution. It is closely related to the larger problem of digital data processing, because the amount of data needed is usually large, and processing it either manually or in analog form is often not feasible. We first consider the span of observation time of the experimental data, which is a fundamental limitation, irrespective of the means of processing the data.

The time span of the data to be analyzed must, of course, be finite; and, as a practical matter, we prefer not to analyze any more data than is necessary to

achieve reasonable results. Remember that since this is a matter of statistical inference, there will always remain some statistical uncertainty in the result. One way to specify the accuracy of the experimentally determined spectrum or autocorrelation function is to say that its variance must be less than a specified value. General accuracy bounds applicable to all processes are not available but there is one special case, the Gaussian process, that is amenable to analysis. We will not give the proof here, but it is shown in a number of references (10,11) that the variance of an experimentally determined autocorrelation function satisfies the inequality

$$\text{Var } V_X(\tau) \leq \frac{4}{T} \int_0^{\infty} R_X^2(\tau) d\tau \quad (2.15.1)$$

where it is assumed that a single sample realization of the process is being analyzed, and

T = time length of the experimental record.

$R_X(\tau)$ = autocorrelation function of the Gaussian process under consideration.

$V_X(\tau)$ = time average of $X_T(t)X_T(t + \tau)$ where $X_T(t)$ is the finite-length sample of $X(t)$. (That is, $V_X(\tau)$ is the experimentally determined autocorrelation function based on a finite record length.)

It should be mentioned that in determining the time average of $X_T(t)X_T(t + \tau)$, we cannot use the whole span of time T , because $X_T(t)$ must be shifted an amount of τ with respect to itself before multiplication. The true extension of $X_T(t)$ beyond the experimental data span is unknown; therefore, we simply omit the nonoverlapped portion in the integration:

$$V_X(\tau) = [\text{time ave. of } X_T(t)X_T(t + \tau)] = \frac{1}{T - \tau} \int_0^{T-\tau} X_T(t)X_T(t + \tau) dt \quad (2.15.2)$$

It will be assumed from this point on that the range of τ being considered is much less than the total data span T , that is, $\tau \ll T$.

We first note that $V_X(\tau)$ is the result of analyzing a single time signal; therefore, $V_X(\tau)$ is itself just a sample function from an ensemble of functions. It is hoped that $V_X(\tau)$ as determined by Eq. (2.15.2) will yield a good estimate of $R_X(\tau)$ and, in order to do so, it should be an unbiased estimator. This can be verified by computing its expectation:

$$\begin{aligned} E[V_X(\tau)] &= E\left[\frac{1}{T - \tau} \int_0^{T-\tau} X_T(t)X_T(t + \tau) dt\right] \\ &= \frac{1}{T - \tau} \int_0^{T-\tau} E[X_T(t)X_T(t + \tau)] dt \\ &= \frac{1}{T - \tau} \int_0^{T-\tau} R_X(\tau) dt = R_X(\tau) \end{aligned} \quad (2.15.3)$$

Thus, $V_X(\tau)$ is an unbiased estimator of $R_X(\tau)$. Also, it can be seen from the equation for $\text{Var } V_X(\tau)$, Eq. (2.15.1), that if the integral of R_X^2 converges (e.g., R_X decreases exponentially with τ), then the variance of $V_X(\tau)$ approaches zero as T becomes large. Thus $V_X(\tau)$ would appear to be a well-behaved estimator of $R_X(\tau)$, that is, $V_X(\tau)$ converges in the mean to $R_X(\tau)$. We will now pursue the estimation accuracy problem further.

Equation (2.15.1) is of little value if the process autocorrelation function is not known. So, at this point, we assume that $X(t)$ is a Gauss-Markov process with an autocorrelation function

$$R_X(\tau) = \sigma^2 e^{-\beta|\tau|} \quad (2.15.4)$$

The σ^2 and β parameters may be difficult to determine in a real-life problem, but we can get at least a rough estimate of the amount of experimental data needed for a given required accuracy. Substituting the assumed Markov autocorrelation function into Eq. (2.15.1) then yields

$$\text{Var}[V_X(\tau)] \leq \frac{2\sigma^4}{\beta T} \quad (2.15.5)$$

We now look at an example illustrating the use of Eq. (2.15.5).

Example 2.14 Let us say that the process being investigated is thought to be a Gauss-Markov process with an estimated time constant ($1/\beta$) of 1 sec. Let us also say that we wish to determine its autocorrelation function within an accuracy of 10 percent, and we want to know the length of experimental data needed. By "accuracy" we mean that the experimentally determined $V(\tau)$ should have a standard deviation less than .1 of the σ^2 of the process, at least for a reasonably small range of τ near zero. Therefore, the ratio of $\text{Var}[V(\tau)]$ to $(\sigma^2)^2$ must be less than .01. Using Eq. (2.15.5), we can write

$$\frac{\text{Var}[V(\tau)]}{\sigma^4} \leq \frac{2}{\beta T}$$

Setting the quantity on the left side equal to .01 and using the equality condition yields

$$T = \frac{1}{(.1)^2} \cdot \frac{2}{\beta} = 200 \text{ sec} \quad (2.15.6)$$

A sketch indicating a typical sample experimental autocorrelation function is shown in Fig. 2.23. Note that 10 percent accuracy is really not an especially demanding requirement, but yet the data required is 200 times the time constant of the process. To put this in more graphic terms, if the process under investigation were random gyro drift with an estimated time constant of 10 hours, 2000 hours of continuous data would be needed to achieve 10 percent accuracy. This could very well be in the same range as the mean time to failure for the gyro. Were we to be more demanding and ask for 1 percent accuracy, about 23 years

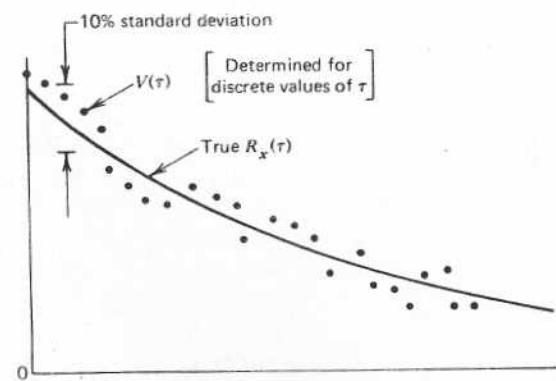


Figure 2.23. Experimental and true autocorrelation functions for Example 2.14.

of data would be required! It can be seen that accurate determination of the autocorrelation function is not a trivial problem in some applications. ■

The main point to be learned from this example is that reliable determination of the autocorrelation function takes considerably more experimental data than one might expect intuitively. The spectral density function is just the Fourier transform of the autocorrelation function, so we might expect a similar accuracy problem in its experimental determination.

As just mentioned, the spectral density function for a given sample signal may be estimated by taking the Fourier transform of the experimentally determined autocorrelation function. This, of course, involves a numerical procedure of some sort because the data describing $V_X(\tau)$ will be in numerical form. The spectral function may also be estimated directly from the periodogram of the sample signal. Recall from Section 2.7 that the average periodogram (the square of the magnitude of the Fourier transform of X_T) is proportional to the spectral density function (for large T). Unfortunately, since we do not usually have the luxury of having a large ensemble of periodograms to average, there are pitfalls in this approach, just as there are in going the autocorrelation route. Nevertheless, modern digital processing methods using Fast Fourier Transform (FFT) techniques have popularized the periodogram approach. Thus it is important to understand its limitations (11).

First, there is the truncation problem. When the time record being analyzed is finite in length, we usually assume that the signal will "jump" abruptly to zero outside the valid data interval. This causes frequency spreading and gives rise to high-frequency components that are not truly representative of the process under consideration, which is assumed to ramble on indefinitely in a continuous manner. An extreme case of this would occur if we were to chop up one long record into many very short records and then average the periodograms of the short records. The individual periodograms, with their predominance of high-frequency components due to the truncation, would not be at all

2.15 DETERMINATION OF AUTOCORRELATION AND SPECTRAL DENSITY FUNCTIONS 107

representative of the spectral content of the original signal; nor would their average! Thus the first rule is that we must have a long time record relative to the typical time variations in the signal. This is true regardless of the method used in analyzing the data. There is, however, a statistical convergence problem that arises as the record length becomes large, and this will now be examined.

In Section 2.7 it was shown that the expectation of the periodogram approaches the spectral density of the process for large T . This is certainly desirable, because we want the periodogram to be an unbiased estimate of the spectral density. It is also of interest to look at the behavior of the variance of the periodogram as T becomes large. Let us denote the periodogram of $X_T(\tau)$ as $M(\omega, T)$, that is,

$$M(\omega, T) = \frac{1}{T} |\mathcal{F}\{X_T(t)\}|^2 \quad (2.15.7)$$

Note that the periodogram is a function of the record length T as well as ω . The variance of $M(\omega, T)$ is

$$\text{Var } M = E(M^2) - [E(M)]^2 \quad (2.15.8)$$

Since we have already found $E(M)$ as given by Eqs. (2.7.8) and (2.7.9), we now need to find $E(M^2)$. Squaring Eq. (2.15.7) leads to

$$E(M^2) = \frac{1}{T^2} E \left[\int_0^T \int_0^T \int_0^T \int_0^T X(t)X(s)X(u)X(v)e^{-j\omega(t-s+u-v)} dt ds du dv \right] \quad (2.15.9)$$

It can be shown that if $X(t)$ is a Gaussian process,*

$$E[X(t)X(s)X(u)X(v)] = R_X(t-s)R_X(u-v) + R_X(t-u)R_X(s-v) + R_X(t-v)R_X(s-u) \quad (2.15.10)$$

Thus, moving the expectation operator inside the integration in Eq. (2.15.9) and using Eq. (2.15.10) leads to

$$\begin{aligned} E(M^2) &= \frac{1}{T^2} \int_0^T \int_0^T \int_0^T \int_0^T [R_X(t-s)R_X(u-v) + R_X(t-u)R_X(s-v) \\ &\quad + R_X(t-v)R_X(s-u)]e^{-j\omega(t-s+u-v)} dt ds du dv \\ &= \frac{1}{T^2} \int_0^T \int_0^T R_X(t-s)e^{-j\omega(t-s)} dt ds \int_0^T \int_0^T R_X(u-v)e^{-j\omega(u-v)} du dv \\ &\quad + \frac{1}{T^2} \int_0^T \int_0^T R_X(t-v)e^{-j\omega(t-v)} dt dv \int_0^T \int_0^T R_X(s-u)e^{-j\omega(s-u)} ds du \\ &\quad + \frac{1}{T^2} \left| \int_0^T \int_0^T R_X(t-u)e^{-j\omega(t+u)} dt du \right|^2 \end{aligned} \quad (2.15.11)$$

* See Problem 2.28.

Next, substituting Eq. (2.7.3) into (2.15.11) leads to

$$E(M^2) = 2[E(M)]^2 + \frac{1}{T^2} \left| \int_0^T \int_0^T R_X(t-u)e^{-j\omega(t+u)} dt du \right|^2 \quad (2.15.12)$$

Therefore

$$\begin{aligned} \text{Var } M &= E(M^2) - [E(M)]^2 \\ &= [E(M)]^2 + \frac{1}{T^2} \left| \int_0^T \int_0^T R_X(t-u)e^{-j\omega(t+u)} dt du \right|^2 \end{aligned} \quad (2.15.13)$$

The second term of Eq. (2.15.13) is nonnegative, so it should be clear that

$$\text{Var } M \geq [E(M)]^2 \quad (2.15.14)$$

But $E(M)$ approaches the spectral function as $T \rightarrow \infty$. Thus, the variance of the periodogram does not go to zero as $T \rightarrow \infty$ (except possibly at those exceptional points where the spectral function is zero). In other words, the periodogram does not converge in the mean as $T \rightarrow \infty$! This is most disturbing, especially in view of the popularity of the periodogram method of spectral determination. The dilemma is summarized in Fig. 2.24. Increasing T will not help reduce the ripples in the individual periodogram. It simply makes M "jump around" faster with ω . This does help, though, with the subsequent averaging that must accompany the spectral analysis. Recall that it is the *average* periodogram that is the measure of the spectral density function. Averaging may not be essential in the analysis of deterministic signals, but it is for random signals. Averaging in both frequency and time is easily accomplished in analog spectrum analyzers by appropriate adjustment of the width of the scanning window and the sweep speed. In digital analyzers, similar averaging over a band of discrete frequencies can be implemented in software. Also, further averaging in time may be accomplished by averaging successive periodograms before displaying the spectrum graphically. In either event, analog or digital, some form of averaging is essential when analyzing noise.

Our treatment of the general problem of autocorrelation and spectral determination from experimental data must be brief. However, the message here should be clear. Treat this problem with respect. It is fraught with subtleties and pitfalls. Engineering literature abounds with reports of shoddy spectral analysis methods and the attendant questionable results. Know your digital signal processing methods and recognize the limitations of the results.

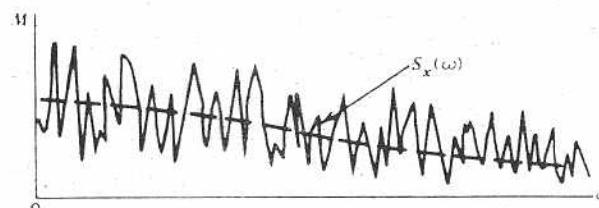


Figure 2.24. Typical periodogram for long record length.

We will pursue the subject of digital spectral analysis further in Section 2.17. But first we digress to present Shannon's sampling theorems, which play an important role in digital signal processing.

2.16 SAMPLING THEOREM

Consider a time function $g(t)$ that is bandlimited, that is,

$$\mathcal{F}[g(t)] = G(\omega) = \begin{cases} \text{Nontrivial}, & |\omega| \leq 2\pi W \\ 0, & |\omega| > 2\pi W \end{cases} \quad (2.16.1)$$

Under the conditions of Eq. (2.16.1), the time function can be written in the form

$$g(t) = \sum_{n=-\infty}^{\infty} g\left(\frac{n}{2W}\right) \frac{\sin(2\pi Wt - n\pi)}{2\pi Wt - n\pi} \quad (2.16.2)$$

This remarkable theorem is due to C. E. Shannon (12,13), and it has special significance when dealing with bandlimited noise.* The theorem says that if one were to specify an infinite sequence of sample values $\dots, g_1, g_2, g_3, \dots$, uniformly spaced $1/2W$ sec apart as shown in Fig. 2.25, then there would be one and only one bandlimited function that would go through all the sample values. In other words, specifying the signal sample values and requiring $g(t)$ to be bandlimited indirectly specifies the signal between the sample points as well. The sampling rate of $2W$ Hz is known as the *Nyquist rate*. This represents the minimum sampling rate needed to preserve all the information content in the continuous signal. If we sample $g(t)$ at less than the Nyquist rate, some information will be lost, and the original signal cannot be exactly reconstructed on the basis of the sequence of samples. Sampling at a rate higher than the Nyquist rate is not necessary, but it does no harm because this simply extends the allowable range of signal frequencies beyond W Hz. Certainly, a signal lying within bandwidth W also lies within a bandwidth greater than W .

In describing a stationary random process that is bandlimited, it can be seen that we only need to consider the statistical properties of samples taken at the Nyquist rate of $2W$ Hz. This simplifies the process description considerably. If we add the further requirement that the process is Gaussian and white within the bandwidth W , then the joint probability density for the samples may be written as a simple product of single-variate normal density functions. This simplification is frequently used in noise analysis in order to make the problem mathematically tractable.

* The basic concept of sampling at twice the highest signal frequency is usually attributed to H. Nyquist (14). However, the explicit form of the sampling theorem given by Eq. (2.16.2) and its associated signal bandwidth restriction was first introduced into communication theory by C. E. Shannon. You may wish to refer to Shannon (13) or Black (15) for further comments on the history of sampling theory.

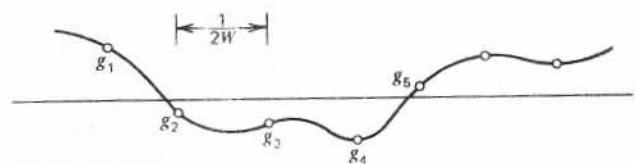


Figure 2.25. Samples of bandlimited signal $g(t)$.

Since there is symmetry in the direct and inverse Fourier transforms, we would expect there to be a corresponding sampling theorem in the frequency domain. It may be stated as follows. Consider the time function $g(t)$ to be time limited, that is, nontrivial over a span of time T and zero outside this interval; then its Fourier transform $G(\omega)$ may be written as

$$G(\omega) = \sum_{n=-\infty}^{\infty} G\left(\frac{2\pi n}{T}\right) \frac{\sin\left(\frac{\omega T}{2} - n\pi\right)}{\frac{\omega T}{2} - n\pi} \quad (2.16.3)$$

All of the previous comments relative to time domain sampling have their corresponding frequency-domain counterparts.

Frequently, it is useful to consider time functions that are limited in both time and frequency. Strictly speaking, this is not possible, but it is a useful approximation. This being the case, the time function can be uniquely represented by $2WT$ samples. These may be specified either in the time or frequency domain.

Sampling theorems have also been worked out for the nonbaseband case (15,16,19). These are somewhat more involved than the baseband theorems and will not be given here.

2.17 DISCRETE FOURIER TRANSFORM AND FAST FOURIER TRANSFORM

The subject of digital signal processing has received considerable attention in the past few decades, and it is only natural that this would occur concurrently with the advancement of computer technology. Whole books have been devoted to the subject (11,17,18), so that we cannot expect to do the matter justice in one short section. We can, however, present a brief overview in order to place digital spectral analysis in proper perspective. We will then proceed on to the main subject of this book, namely, filter analysis.

Modern computer technology has made it possible to perform an efficient digital version of the Fourier transform. Thus, nearly all spectral analysis is

now done using the direct periodogram approach rather than the more round-about approach via the autocorrelation function. The sampling theorems presented in Section 2.16 dictate some constraints in the choice of sampling rates and the total amount of data analyzed in any one batch. Since these constraints play an important role in digital signal processing, we will examine their consequences in some detail.

In spectral analysis, we usually have at least a rough idea as to the bandwidth of the signal to be analyzed; therefore, let us say this is approximately 0 to W Hz. The sampling theorem, Eq. (2.16.2), says that the sampling rate should be $2W$ samples/sec or, equivalently, the sample spacing should be $1/2W$ sec. Let us further say that we wish to analyze N samples in one batch where N is yet to be determined. The total time span of the samples would then be

$$\text{Total time span of data} = T = \frac{N}{2W} \quad (2.17.1)$$

The frequency-domain sampling theorem, Eq. (2.16.3), states that our truncated time signal could be represented in the spectral domain with discrete samples spaced $1/T$ or $2W/N$ apart. That is, we have N samples uniformly spaced in the time domain and $N/2$ corresponding samples spaced uniformly from 0 to W Hz in the frequency domain. Since each spectral sample is a complex number, the number of degrees of freedom is N in either the time or frequency domain. This one-to-one correspondence of scalar elements in the two domains suggests a transform-pair relationship, and this will now be formalized.

As a matter of notation let the truncated time signal be $g(t)$ and its Fourier transform be

$$G(j\omega) = \int_0^T g(t)e^{-j\omega t} dt \quad (2.17.2)$$

Consider next a discrete approximation for $G(j\omega)$ as follows:

$$G(jn2\pi\Delta f) = \sum_{k=0}^{N-1} g_k e^{-jn2\pi \Delta f k \Delta T} \Delta T \quad (2.17.3)$$

where

g_k = sequence of N time samples of $g(t)$

$\Delta T = \frac{1}{2W}$ = sample spacing in time domain

$\Delta f = \frac{2W}{N}$ = sample spacing in frequency domain (in hertz)

We note now that

$$\Delta f \Delta T = 1/N \quad (2.17.4)$$

Thus, Eq. (2.17.3) can be rewritten in the form

$$G(jn2\pi \Delta f) \Delta f \approx \frac{1}{N} \sum_{k=0}^{N-1} g_k \exp\left(-j \frac{2\pi nk}{N}\right) \quad (2.17.5)$$

Now, given the time sequence g_0, g_1, \dots, g_{N-1} , think of the right side of Eq. (2.17.5) as defining another sequence $\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_{N-1}$ as n is indexed from 0 to $N - 1$. That is, the \mathcal{G}_n sequence is defined as

$$\mathcal{G}_n = \frac{1}{N} \sum_{k=0}^{N-1} g_k \exp\left(-j \frac{2\pi nk}{N}\right), \quad n = 0, 1, \dots, N - 1 \quad (2.17.6)$$

Note that we do not claim the \mathcal{G}_n 's to be exact samples of $G(j\omega)$, but, hopefully, they will be reasonable approximations thereof.

The \mathcal{G}_n sequence as defined by Eq. (2.17.6) exhibits certain symmetry that is worth noting. First, if we extend the index n beyond $N - 1$, we simply get a periodic extension of the sequence:

$$\begin{aligned} \mathcal{G}_N &= \mathcal{G}_0 \\ \mathcal{G}_{N+1} &= \mathcal{G}_1 \\ &\vdots \\ &\text{etc.} \end{aligned} \quad (2.17.7)$$

Also, there is symmetry about the midpoint of the sequence in that

$$\begin{aligned} \mathcal{G}_{N-1} &= \mathcal{G}_1^* \\ \mathcal{G}_{N-2} &= \mathcal{G}_2^* \\ &\vdots \\ &\text{etc.} \end{aligned} \quad (2.17.8)$$

In other words, half of the defined \mathcal{G}_n 's are complex conjugates of the other half and are thus redundant (see Problem 2.29). This is as expected; there can only be N degrees of freedom in the frequency domain, just as in the time domain. Figure 2.26 summarizes the symmetry properties of the \mathcal{G}_n sequence.

Once the \mathcal{G}_n sequence is defined as per Eq. (2.17.6), it can be shown that an exact inverse relationship exists as follows (17,18):

$$g_k = \sum_{n=0}^{N-1} \mathcal{G}_n \exp\left(j \frac{2\pi nk}{N}\right), \quad k = 0, 1, \dots, N - 1 \quad (2.17.9)$$

The \mathcal{G}_n and g_k sequences then form a transform pair in the usual sense; that is, given one, we can find the other, and vice versa.* As might be expected, the

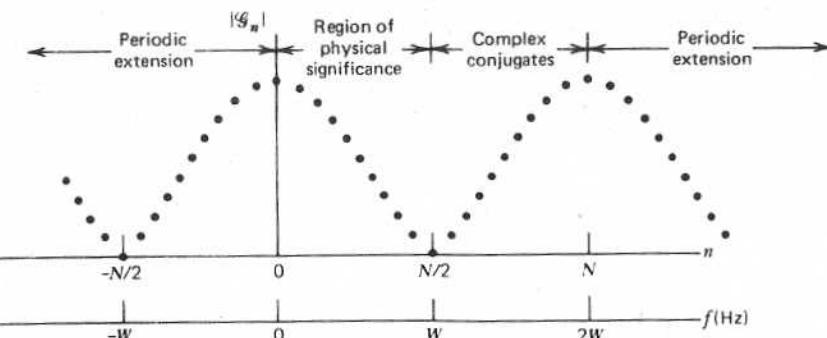


Figure 2.26. Discrete Fourier transform magnitudes and their significance.

\mathcal{G}_n sequence is called the *discrete Fourier transform* of g_k and, of course, g_k is the discrete inverse transform of \mathcal{G}_n . It should be emphasized that there is no approximation involved in the discrete transform pair relationship. The approximation comes in the *interpretation* of \mathcal{G}_n as samples of the signal spectrum. This is a relatively complicated matter and the references cited (11,15,16) may be consulted for further details. It suffices here to say that this has been studied extensively and, with appropriate cautions and proper weighting of the time data, the discrete Fourier transform can provide meaningful results.

We might also note that all the preceding remarks about digital signal processing apply to deterministic as well as random signals. In analyzing deterministic signals, we usually compute the magnitude of the discrete Fourier transform and call this the signal spectrum. In the random signal case, usually the square of the magnitudes of the \mathcal{G}_n terms are formed, and this sequence (i.e., $|\mathcal{G}_0|^2, |\mathcal{G}_1|^2, \dots$) becomes an approximation of the periodogram of the signal being analyzed. The periodogram is, in turn, statistically related to the power spectral density function, which is usually the desired end result of the analysis.

Digital implementation of the discrete Fourier transform is not a trivial matter. High resolution and reliable results in the frequency domain are obtained by making N large. However, if one programs the transform literally as given by Eq. (2.17.6), the number of multiplications required is of the order N^2 . This can easily get out of hand, especially in "on-line" applications. Fortunately, fast, efficient algorithms have been developed for which the number of required multiplications is of the order of $N \log_2 N$ rather than N^2 (17,18). The computational saving is spectacular for large N . For example, let N be $2^{10} = 1024$, which is a modest number of time samples for many applications. Then N^2 would be about 10^6 , whereas $N \log_2 N$ is only about 10^4 . This represents a saving of about a factor of 100 and reflects directly into the time required for the transformation.

All of the fast discrete Fourier transform algorithms require that the number of samples be an integer power 2 (which presents no particular problem).

* Some authors prefer to associate the $1/N$ factor of Eq. (2.17.6) with the inverse relationship

and they all go under the generic name of *Fast Fourier Transform* (FFT). The FFT cannot perform any wondrous, magical tricks on the basic data (as some seem to believe); it is simply an efficient means of implementing the discrete Fourier transform. Thus, all of the cautions that apply to the discrete Fourier transform also apply to the FFT. Because of its efficiency, though, the FFT is used almost universally in both on-line and off-line spectral analysis applications. After all, why do something the hard way when there is an easy way!

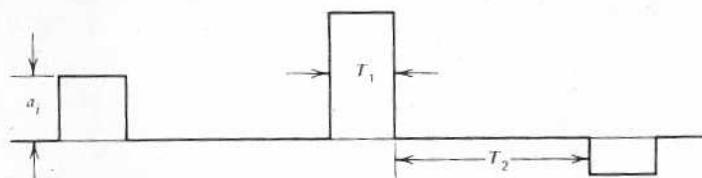
Problems

2.1. Noise measurements at the output of a certain amplifier (with its input shorted) indicate that the rms output voltage due to internal noise is $100\mu\text{V}$. If we assume that the frequency spectrum of the noise is flat from 0 to 10 MHz and zero above 10 MHz, find:

- (a) The spectral density function for the noise.
- (b) The autocorrelation function for the noise.

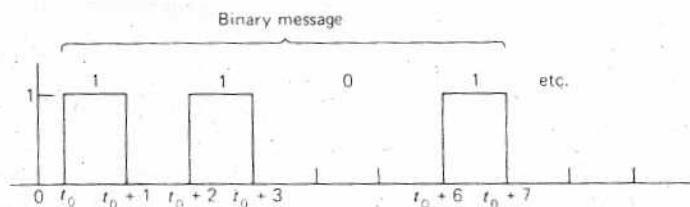
Give proper units for both the spectral density and autocorrelation functions.

2.2. A sketch of a sample realization of a random process $X(t)$ is shown in the figure. The pulse amplitudes a_i are independent samples of a normal random variable with zero mean and variance σ^2 . The time origin is completely random. Find the autocorrelation function for the process.



Problem 2.2

2.3. The waveform shown is an example of a digital-coded waveform. The signal is equally likely to be zero or one in the intervals $(t_0, t_0 + 1)$, $(t_0 + 2, t_0 + 3)$, etc., and it is always zero in the "in between" intervals $(t_0 + 1, t_0 + 2)$, $(t_0 + 3, t_0 + 4)$, etc. The initial switching time t_0 is random and uniformly distributed between zero and one. The presence or absence of a pulse in the "pulse possible" intervals is the code for a binary digit. There is no statistical correlation among any of the bits of the message. Find the autocorrelation and spectral density functions for this process.



Problem 2.3

PROBLEMS 115
2.4. Find the autocorrelation function corresponding to the spectral density function

$$S(j\omega) = \delta(\omega) + \frac{1}{2}\delta(\omega - \omega_0) + \frac{1}{2}\delta(\omega + \omega_0) + e^{-2|\omega|}$$

2.5. A stationary Gaussian random process $X(t)$ has an autocorrelation function of the form

$$R_X(\tau) = 4e^{-|\tau|}$$

What fraction of the time will $|X(t)|$ exceed four units?

2.6. A random process $X(t)$ is generated according to the following rules:

- (a) The waveform is generated with a sample-and-hold arrangement where the "hold" interval is 1 sec.
- (b) The successive amplitudes for each 1-sec interval are independent samples taken from a zero-mean normal distribution with a variance of σ^2 .
- (c) The first switching time is a random variable with uniform distribution from 0 to 1 (i.e., the time origin is random).

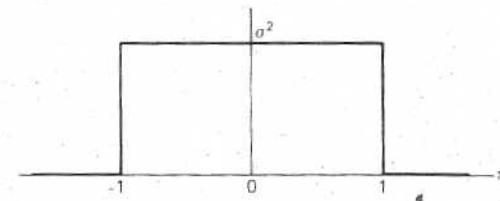
Let X_1 denote $X(t_1)$ and X_2 denote $X(t_1 + \tau)$.

- (a) Find the joint probability density function $f_{X_1 X_2}(x_1, x_2)$.
- (b) Is this process a Gaussian process?

2.7. Find the autocorrelation function for the process described in Problem 2.6 using the expectation formula

$$R_X(\tau) = E[X_1 X_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X_1 X_2}(x_1, x_2) dx_1 dx_2$$

2.8. It is suggested that a certain real process has an autocorrelation function as shown in the figure. Is this possible? Justify your answer. (Hint: Calculate the spectral density function and see if it is plausible.)



Problem 2.8

2.9. Consider the random process $X(t) = 2 \sin \omega t$ where ω is a random variable with uniform distribution between $\omega = 2$ and $\omega = 6$. Is the process (a) stationary, (b) ergodic, and (c) deterministic or nondeterministic?

2.10. The input to an ideal rectifier (unity forward gain, zero reverse gain) is a stationary Gaussian process.

- (a) Is the output stationary?
- (b) Is the output a Gaussian process?

Give a brief justification for both answers.

2.11. A random process $X(t)$ has sample realizations of the form

$$X(t) = at + Y$$

where a is a known constant and Y is a random variable whose distribution is $N(0, \sigma^2)$. Is the process (a) stationary, and (b) ergodic? Justify your answers.

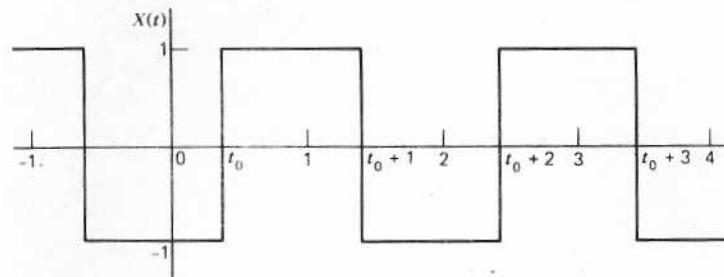
2.12. What is the autocorrelation function for $X(t)$ of Problem 2.11?

2.13. A sample realization of a random process $X(t)$ is shown in the figure. The time t_0 when the transition from the -1 state to the $+1$ takes place is a random variable that is uniformly distributed between 0 and 2 units.

(a) Is the process stationary?

(b) Is the process deterministic or nondeterministic?

(c) Find the autocorrelation function and spectral density function for the process.



Problem 2.13

2.14. A common autocorrelation function encountered in physical problems is

$$R(\tau) = \sigma^2 e^{-\beta|\tau|} \cos \omega_0 \tau$$

(a) Find the corresponding spectral density function.

(b) $R(\tau)$ will be recognized as a damped cosine function. Sketch both the autocorrelation and spectral density functions for the lightly damped case.

2.15. Show that a Gauss-Markov process described by the autocorrelation function

$$R(\tau) = \sigma^2 e^{-\beta|\tau|}$$

becomes Gaussian white noise if we let $\beta \rightarrow \infty$ and $\sigma^2 \rightarrow \infty$ in such a way that the area under the autocorrelation-function curve remains constant in the limiting process.

2.16. A stationary random process $X(t)$ has a spectral density function of the form

$$\dot{S}_X(\omega) = \frac{6\omega^2 + 12}{(\omega^2 + 4)(\omega^2 + 1)}$$

What is the mean square value of $X(t)$?

[Hint: $S_X(\omega)$ may be resolved into a sum of two terms of the form: $A/(\omega^2 + 4) + B/(\omega^2 + 1)$. Each term may then be integrated using standard integral tables.]

2.17. The stationary process $X(t)$ has an autocorrelation function of the form

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

Another process $Y(t)$ is related to $X(t)$ by the deterministic equation

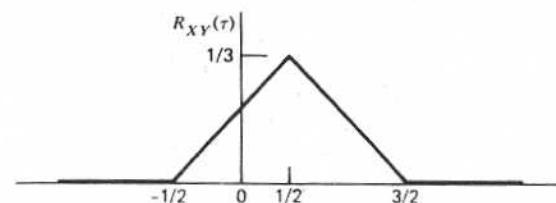
$$Y(t) = aX(t) + b$$

where a and b are known constants.

(a) What is the autocorrelation function for $Y(t)$?

(b) What is the crosscorrelation function $R_{XY}(\tau)$?

2.18. The crosscorrelation function $R_{XY}(\tau)$ for Example 2.8 is sketched below. What is the corresponding cross spectral density function?



Problem 2.18

2.19. The random telegraph wave is described in Section 2.11. Let this process be the input to an ideal rectifier (unity forward gain, zero reverse gain).

(a) What is the autocorrelation function of the output?

(b) What is the crosscorrelation function $R_{XY}(\tau)$ (X is input, Y is output)?

2.20. Two deterministic random processes are defined by

$$X(t) = A \sin (\omega t + \theta)$$

$$Y(t) = B \sin (\omega t + \theta)$$

where θ is a random variable with uniform distribution between 0 and 2π , and ω is a known constant. The A and B coefficients are both normal random variables $N(0, \sigma^2)$ and are correlated with a correlation coefficient ρ . What is the crosscorrelation function $R_{XY}(\tau)$? (Assume A and B are independent of θ .)

2.21. The discrete random walk process is discussed in Section 2.13. Assume each step is of length l and that the steps are independent and equally likely to be positive or negative. Show that the variance of the total distance D traveled in N steps is given by

$$\text{Var } D = l^2 N$$

[Hint: First write D as the sum $l_1 + l_2 + \dots + l_N$ and note that l_1, l_2, \dots, l_N are independent random variables. Then form $E(D)$ and $E(D^2)$ and compute $\text{Var } D$ as $E(D^2) - [E(D)]^2$.]

2.22. Let the process $Z(t)$ be the product of two independent stationary processes $X(t)$ and $Y(t)$. Show that the spectral density function for $Z(t)$ is given by (in the s domain)

$$S_Z(s) = \frac{1}{2\pi j} \int_{-\infty}^{j\infty} S_X(w)S_Y(s-w) dw$$

[Hint: First show that $R_Z(\tau) = R_X(\tau)R_Y(\tau)$.]

2.23. The spectral density function for the stationary process $X(t)$ is

$$S_X(j\omega) = \frac{1}{(1 + \omega^2)^2}$$

Find the autocorrelation function for $X(t)$.

- 2.24.** A stationary process $X(t)$ is Gaussian and has an autocorrelation function of the form

$$R_X(\tau) = 4e^{-|\tau|}$$

Let the random variable X_1 denote $X(t_1)$ and X_2 denote $X(t_1 + 1)$. Write the expression for the joint probability density function $f_{X_1 X_2}(x_1, x_2)$.

- 2.25.** The Global Positioning System (GPS) is a satellite navigation system that is scheduled to be fully operational in the late 1980s (8,9). The signals transmitted by the satellites provide timing information that enables the receiver to determine its position and velocity in an earth-fixed reference frame. One of the signals provided by each of the system satellites is of the form

$$s(t) = C(t) \sin 2\pi f_c t$$

where

$$f_c = \text{carrier frequency} = 1575.42 \text{ MHz}$$

and $C(t)$ is a pseudorandom binary (+1 or -1) code signal similar to the one described in Example 2.13. The code repeats itself after 1023 bits, and the repetition period is exactly 1 ms. Thus, the "code rate" is 1.023 MHz. (Note: There are exactly 1540 cycles of the carrier per bit of the pseudorandom code.)

- (a) Sketch the time autocorrelation function for $C(t)$.
- (b) Sketch the spectrum for $C(t)$. Since the spectrum will consist of discrete lines, let the line lengths in the sketch indicate the harmonic magnitudes.
- (c) Sketch the spectrum for the radiofrequency signal $s(t)$.

(Actually, the transmittal signal has an additional low-frequency modulation signal that contains the navigational message. This has the effect of spreading the spectral lines slightly. This may be ignored in this problem.)

- 2.26.** We wish to determine the autocorrelation function a random signal empirically from a single time record. Let us say we have good reason to believe the process is ergodic and at least approximately Gaussian and, furthermore, that the autocorrelation function of the process decays exponentially with a time constant no greater than 10 sec. Estimate the record length needed to achieve 5 percent accuracy in the determination of the autocorrelation function. (By 5 percent accuracy, assume we mean that for any τ , the standard deviation of the experimentally determined autocorrelation function will not be more than 5 percent of the maximum value of the true autocorrelation function.)

- 2.27.** In Problem 2.26 the signal is not known to be truly bandlimited, but it is reasonable to assume that essentially all the signal energy will lie between 0 and .1 Hz. Let us assume .1 Hz to be the signal bandwidth, and let us say we wish to sample the signal at the Nyquist rate.

- (a) How many discrete samples would be required to describe the record length of Problem 2.26?
- (b) Suppose that we wish to compute the discrete Fourier transform of the finite-time signal using the Fast Fourier Transform. This requires that the number of samples N be an integer power of 2. What should N be in this case?
- (c) The value of N in part (b) should work out to be greater than that found in part (a). In order to achieve the appropriate N for the FFT algorithm, would we be better

off (in terms of accuracy) to increase the sampling rate for the length of time computed in Problem 2.26, or should we keep the sampling rate at .2 Hz and increase the time length of the record accordingly? Presumably, the computational effort would be the same either way.

- 2.28.** Let X_1, X_2, X_3, X_4 be zero-mean Gaussian random variables. Show that

$$E(X_1 X_2 X_3 X_4) = E(X_1 X_2) E(X_3 X_4) + E(X_1 X_3) E(X_2 X_4) + E(X_1 X_4) E(X_2 X_3) \quad (\text{P2.28.1})$$

(Hint: The characteristic function was discussed briefly in Section 1.8. The multivariate version of the characteristic function is useful here. Let $\psi(\omega_1, \omega_2, \dots, \omega_n)$ be the multidimensional Fourier transform of $f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n)$ (but with the signs reversed on $\omega_1, \omega_2, \dots, \omega_n$). Then it can be readily verified that

$$\left(-j \right)^n \frac{\partial^n \psi(\omega_1, \omega_2, \dots, \omega_n)}{\partial \omega_1 \partial \omega_2 \dots \partial \omega_n} \Bigg|_{\substack{\omega_1=0 \\ \omega_2=0 \\ \dots \\ \omega_n=0}} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_1 x_2 \dots x_n \dots$$

$$x_n f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

$$= E(X_1, X_2, \dots, X_n) \quad (\text{P2.28.2})$$

The characteristic function for a zero-mean, vector Gaussian random variable \mathbf{X} is

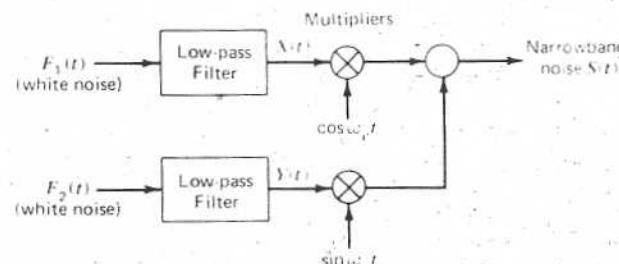
$$\psi(\omega) = e^{-\frac{1}{2} \omega^T \mathbf{C}_X \omega} \quad (\text{P2.28.3})$$

where \mathbf{C}_X is the covariance matrix for \mathbf{X} . This, along with Eq. (P2.28.2), may now be used to justify the original statement given by Eq. (P2.28.1).

- 2.29.** It was mentioned in Section 2.17 that half of the discrete Fourier transform elements are complex conjugates of the other half. This statement deserves closer scrutiny because N may be either odd or even insofar as the basic discrete transform is concerned (and not its efficient FFT implementation).

- (a) For the case where N is even, show that \mathcal{G}_0 and $\mathcal{G}_{N/2}$ are both real, and thus the total number of nonredundant scalar elements in the frequency domain is N , just as in the time domain.
- (b) For the case where N is odd, show that only \mathcal{G}_0 is constrained to be real, and thus the count of nonredundant scalar elements is N , just as in the previous case.

- 2.30.** The accompanying figure shows a means of generating narrowband noise from two independent baseband noise sources. (See Section 2.12.) The bandwidth of the resulting narrow band noise is controlled by the cutoff frequency of the low-pass filters.



Problem 2.30

which are assumed to have identical characteristics. Assume that $F_1(t)$ and $F_2(t)$ are independent white Gaussian noise processes with similar spectral amplitudes. The resulting noise processes after low-pass filtering will then have identical autocorrelation functions that will be denoted $R_X(\tau)$.

- Show that the narrowband noise $S(t)$ is a stationary Gaussian random process whose autocorrelation function is

$$R_S(\tau) = R_X(\tau) \cos \omega_c \tau$$

- Also show that both the in-phase and quadrature channels are needed to produce stationary narrow band noise. (That is, if either of the $\sin \omega_c t$ or $\cos \omega_c t$ multiplying operations is omitted, the resultant output will not be strictly stationary.)

References Cited in Chapter 2

- E. Wong, *Stochastic Processes in Information and Dynamical Systems*, New York: McGraw-Hill, 1971.
- A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, New York: Academic Press, 1970.
- N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Cambridge, Mass.: MIT Press, and New York: Wiley; 1949.
- S. O. Rice, "Mathematical Analysis of Noise," *Bell System Tech. J.*, 23, 282-332 (1944); 24, 46-256 (1945).
- W. B. Davenport, Jr., and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.
- A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965.
- R. C. Dixon, *Spread Spectrum Systems*, New York: Wiley, 1976.
- R. P. Denaro, "Navstar: The All-Purpose Satellite," *IEEE Spectrum*, 18:5, 35-40 (May 1981).
- R. J. Milliken and C. J. Zoller, "Principle of Operation of NAVSTAR and System Characteristics," *Navigation*, Jour. of the Inst. of Navigation, 25:2 (Special Issue on GPS), 95-106. (Summer 1978).
- J. H. Laning, Jr. and R. H. Battin, *Random Processes in Automatic Control*, New York: McGraw-Hill, 1956.
- J. S. Bendat, and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, New York: Wiley-Interscience, 1971.
- C. E. Shannon, *The Mathematical Theory of Communication*, "Bell System Tech. J.", July and October 1948. (Later reprinted in book form by the University of Illinois Press, 1949.)
- C. E. Shannon, "Communication in the Presence of Noise," *Proc. of the Inst. of Radio Engr.*, v. 37, No. 1, January 1949, pp. 10-21.
- H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Trans. of the Am. Inst. of Elect. Engr.*, v. 47, April 1928, pp. 617-644.
- H. S. Black, *Modulation Theory*, New York: D. Van Nostrand Co., 1950.
- S. Goldman, *Information Theory*, Englewood Cliffs, N.J.: Prentice-Hall, 1953.

- A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- D. Childers and A. Durling, *Digital Filtering and Signal Processing*, St. Paul: West Publishing Co., 1975.
- K. S. Shanmugam, *Digital and Analog Communication Systems*, New York: Wiley, 1979.

Additional References on Random Processes

- P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles*, New York: McGraw-Hill, 1980.
- H. J. Larson and B. O. Shubert, *Probabilistic Models in Engineering Sciences (Vols. 1 and 2)*, New York: Wiley, 1979.
- G. R. Cooper and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, New York: Holt, Rinehart, and Winston, 1971.
- J. L. Melsa and A. P. Sage, *An Introduction to Probability and Stochastic Processes*, Englewood Cliffs, N.J.: Prentice-Hall, 1973.
- A. M. Breipohl, *Probabilistic Systems Analysis*, New York: Wiley, 1970.

CHAPTER 3

Response of Linear Systems to Random Inputs

The central problem of linear systems analysis is: Given the input, what is the output? In the deterministic case, we usually seek an explicit expression for the response or output. In the random-input problem no such explicit expression is possible, except for the special case where the input is a so-called "deterministic" random process (and not always in this case). Usually, in random-input problems, we must settle for a considerably less-complete description of the output than we get for corresponding deterministic problems. In the case of random processes the most convenient descriptors to work with are autocorrelation function, spectral density function, and mean-square value. We now examine the input-output relationships of linear systems in these terms.

3.1 INTRODUCTION: THE ANALYSIS PROBLEM

In any system satisfying a set of linear differential equations, the solution may be written as a superposition of an initial-condition part and another part due to the driving or forcing functions. Both the initial conditions and forcing functions may be random; and, if so, the resultant response is a random process. We direct our attention here to such situations, and it will be assumed that the reader has at least an elementary acquaintance with deterministic methods of linear system analysis (1,2,3).

With reference to Fig. 3.1, the analysis problem may be simply stated: Given the initial conditions and the input and the system's dynamical characteristics [i.e., $G(s)$ in Fig. 3.1], what is the output? Of course, in the stochastic problem, the input and output will have to be described in probabilistic terms.

We need to digress here for a moment and discuss a notational matter. In Chapters 1 and 2 we were careful to use uppercase symbols to denote random variables and lowercase symbols for the corresponding arguments of their probability density functions. This is the custom in most current books on probability. There is, however, a long tradition in engineering books on automatic control and linear systems analysis of using lowercase for time functions and uppercase for the corresponding Laplace or Fourier transforms. So, we are confronted with notational conflict. We will resolve this in favor of the traditional linear analysis notation, and from this point on we will use lowercase

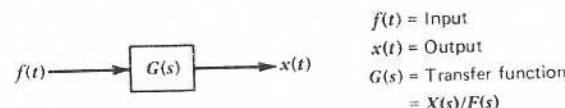


Figure 3.1. Block diagram for elementary analysis problem.

symbols for time signals—either random or deterministic—and uppercase for their transforms. This seems to be the lesser of the two evils. The reader will simply have to interpret the meaning of symbols such as $x(t)$, $f(t)$, etc., within the context of the subject matter under discussion. This usually presents no problem. For example, with reference to Fig. 3.1, $g(t)$ would mean inverse transform of $G(s)$, and it clearly is a deterministic time function. On the other hand, the input and output, $f(t)$ and $x(t)$, will usually be random processes in the subsequent material.

Generally, analysis problems can be divided into two categories:

1. Stationary (steady-state) analysis. Here the input is assumed to be time stationary, and the system is assumed to have fixed parameters with a stable transfer function. This leads to a stationary output, provided the input has been present for a long period of time relative to the system time constants.
2. Nonstationary (transient) analysis. Here we usually consider the driving function as being applied at $t = 0$, and the system may be initially at rest or have nontrivial initial conditions. The response in this case is usually nonstationary. We note that analysis of unstable systems falls into this category, because no steady-state (stationary) condition will exist.

The similarity between these two categories and the corresponding ones in deterministic analysis should be apparent. Just as in circuit analysis, we would expect the transient solution to lead to the steady-state response as $t \rightarrow \infty$. However, if we are only interested in the stationary result, this is getting at the solution the "hard way." Much simpler methods are available for the stationary solution, and these will now be considered.

3.2 STATIONARY (STEADY-STATE) ANALYSIS

We assume in Fig. 3.1 that $G(s)$ represents a stable, fixed-parameter system and that the input is covariance (wide-sense) stationary with a known spectral function. In deterministic analysis, we know that if the input is Fourier transformable, the input spectrum is simply modified by $G(j\omega)$ in going through the filter. In the random process case, one interpretation of the spectral function is

that it is proportional to the magnitude of the *square* of the Fourier transform. Thus the equation relating the input and output spectral functions is*

$$S_x(s) = G(s)G(-s)S_f(s) \quad (3.2.1)$$

Note that Eq. (3.2.1) is written in the s domain where the imaginary axis has the meaning of real angular frequency ω . If you prefer to write Eq. (3.2.1) in terms of ω , just replace s with $j\omega$. Equation (3.2.1) then becomes

$$\begin{aligned} S_x(j\omega) &= G(j\omega)G(-j\omega)S_f(j\omega) \\ &= |G(j\omega)|^2 S_f(j\omega) \end{aligned} \quad (3.2.2)$$

Because of the special properties of spectral functions, both sides of Eq. (3.2.2) work out to be real functions of ω . Also note that the autocorrelation function of the output can be obtained as the inverse Fourier transform of $S_x(s)$. Two examples will now illustrate the use of Eq. (3.2.1).

Example 3.1 Consider a first-order low-pass filter with unity white noise as the input. With reference to Fig. 3.1, then

$$S_f(s) = 1$$

$$G(s) = \frac{1}{1 + Ts}$$

where T is the time constant of the filter. The output spectral function is then

$$\begin{aligned} S_x(s) &= \frac{1}{1 + Ts} \cdot \frac{1}{1 + T(-s)} \cdot 1 \\ &= \frac{(1/T)^2}{-s^2 + (1/T)^2} \end{aligned}$$

Or, in terms of real frequency ω ,

$$S_x(j\omega) = \frac{(1/T)^2}{\omega^2 + (1/T)^2}$$

This is sketched as a function of ω in Fig. 3.2. As would be expected, most of the spectral content is concentrated at low frequencies and then it gradually diminishes as $\omega \rightarrow \infty$.

It is also of interest to compute the mean-square value of the output. It is given by Eq. (2.7.17).

$$E(x^2) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{1}{1 + Ts} \cdot \frac{1}{1 + T(-s)} ds \quad (3.2.3)$$

The integral of Eq. (3.2.3) is easily evaluated in this case by substituting $j\omega$ for s and using a standard table of integrals. This leads to

$$E(x^2) = 1/2T$$

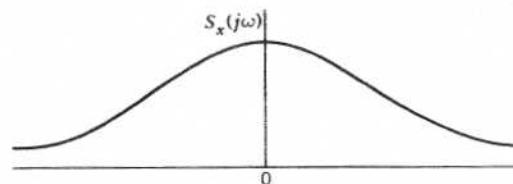


Figure 3.2. Spectral function for low-pass filter output with white noise input.

The "standard" table-of-integrals approach is of limited value, though, as will be seen in the next example. ■

Example 3.2 Consider the input process to have an exponential autocorrelation function and the filter to be the same as in the previous example. Then

$$R_f(\tau) = \sigma^2 e^{-\beta|\tau|}$$

$$G(s) = \frac{1}{1 + Ts}$$

First we transform R_f to obtain the input spectral function.

$$\mathcal{F}[R_f(\tau)] = \frac{2\sigma^2\beta}{-s^2 + \beta^2}$$

The output spectral function is then

$$S_x(s) = \frac{2\sigma^2\beta}{-s^2 + \beta^2} \cdot \frac{(1/T)}{(s + 1/T)} \cdot \frac{(1/T)}{(-s + 1/T)} \quad (3.2.4)$$

Now, if we wish to find $E(x^2)$ in this case, it will involve integrating a function that is fourth order in the denominator and most tables of integrals will be of no help. We note, though, that the input spectral function can be factored and the terms of eq. (3.2.4) can be rearranged as follows:

$$S_x(s) = \left[\frac{\sqrt{2\sigma^2\beta}}{(s + \beta)} \cdot \frac{(1/T)}{(s + 1/T)} \right] \left[\frac{\sqrt{2\sigma^2\beta}}{(-s + \beta)} \cdot \frac{(1/T)}{(-s + 1/T)} \right] \quad (3.2.5)$$

The first term has all its poles and zeros in the left half plane, and the second term has mirror-image poles and zeros in the right half plane. This regrouping of terms is known as *spectral factorization* and can always be done if the spectral function is rational in form. (that is, if it can be written as a ratio of polynomials in even powers of s .)

Since special tables of integrals have been worked out for integrating complex functions of the type given by Eq. (3.2.5), we defer evaluating $E(x^2)$ until these have been presented in the next section. ■

*See Problem 3.1 for a formal justification of Eq. (3.2.1).

3.3 INTEGRAL TABLES FOR COMPUTING MEAN-SQUARE VALUE

In linear analysis problems, the spectral function can often be written as a ratio of polynomials in s^2 . If this is the case, spectral factorization can be used to write the function in the form

$$S_x(s) = \frac{c(s)}{d(s)} \cdot \frac{c(-s)}{d(-s)} \quad (3.3.1)$$

where $c(s)/d(s)$ has all its poles and zeros in the left half plane and $c(-s)/d(-s)$ has mirror-image poles and zeros in the right half plane. No roots of $d(s)$ are permitted on the imaginary axis. The mean-square value of x can now be written as

$$E(x^2) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{c(s)c(-s)}{d(s)d(-s)} ds \quad (3.3.2)$$

R. S. Phillips (4) was the first to prepare a table of integrals for definite integrals of the type given by Eq. (3.3.2). His table has since been repeated in many texts with a variety of minor modifications (5,6). An abbreviated table in terms of the complex s domain follows.

Table 3.1 Table of Integrals

$I_n = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{c(s)c(-s)}{d(s)d(-s)} ds$	(3.3.3)
$c(s) = c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \dots + c_0$	
$d(s) = d_ns^n + d_{n-1}s^{n-1} + \dots + d_0$	
$I_1 = \frac{c_0^2}{2d_0d_1}$	
$I_2 = \frac{c_1^2d_0 + c_0^2d_2}{2d_0d_1d_2}$	
$I_3 = \frac{c_2^2d_1 + (c_1^2 - 2c_0c_2)d_0d_3 + c_0^2d_2d_1}{2d_0d_1(d_1d_2 - d_0d_3)}$	
$I_4 = \frac{c_3^2(-d_0^2d_3 + d_0d_1d_2) + (c_2^2 - 2c_1c_3)d_0d_1d_4 + (c_1^2 - 2c_0c_2)d_0d_3d_4 + c_0^2(-d_1d_2^2 + d_2d_3d_4)}{2d_0d_1(-d_0d_3^2 - d_1^2d_4 + d_1d_2d_3)}$	

An example will now illustrate the use of Table 3.1.

Example 3.3 The solution in Example 2 was brought to the point where the spectral function had been written in the form

$$S_x(s) = \left[\frac{\sqrt{2\sigma^2\beta} \cdot 1/T}{(s + \beta)(s + 1/T)} \right] \left[\frac{\sqrt{2\sigma^2\beta} \cdot 1/T}{(-s + \beta)(-s + 1/T)} \right] \quad (3.3.4)$$

Clearly, S_x has been factored properly with its poles separated into left-half and right-half plane parts. The mean-square value of x is given by

$$E(x^2) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} S_x(s) ds \quad (3.3.5)$$

Comparing the form of $S_x(s)$ in Eq. (3.3.4) with the standard form given in Eq. (3.3.3), we see that

$$c(s) = \frac{\sqrt{2\sigma^2\beta}}{T}$$

$$d(s) = s^2 + (\beta + 1/T)s + \beta/T$$

Thus we can use the I_2 integral of Table 3.1. The coefficients for this case are

$$c_1 = 0 \quad d_2 = 1$$

$$c_0 = \frac{\sqrt{2\sigma^2\beta}}{T} \quad d_1 = (\beta + 1/T)$$

$$d_0 = \beta/T$$

and $E(x^2)$ is then

$$E(x^2) = \frac{c_0^2}{2d_0d_1} = \frac{2\sigma^2\beta/T^2}{2(\beta/T)(\beta + 1/T)} = \frac{\sigma^2}{1 + \beta T}$$

3.4 PURE WHITE NOISE AND BANDLIMITED SYSTEMS

We are now in a position to demonstrate the validity of using the pure white-noise model in certain problems, even though white noise has infinite variance. This will be done by posing two hypothetical mean-square analysis problems:

1. Consider a simple first-order low-pass filter with bandlimited white noise as the input. Specifically, with reference to Fig. 3.1, let

$$S_f(j\omega) = \begin{cases} A, & |\omega| \leq \omega_c \\ 0, & |\omega| > \omega_c \end{cases} \quad (3.4.1)$$

$$G(s) = \frac{1}{1 + Ts} \quad (3.4.2)$$

2. Consider the same low-pass filter as in problem 1, but with pure white noise as the input:

$$S_f(j\omega) = A, \quad (\text{for all } \omega) \quad (3.4.3)$$

$$G(s) = \frac{1}{1 + Ts} \quad (3.4.4)$$

Certainly problem 1 is physically plausible because bandlimited white noise has finite variance. Conversely, problem 2 is not because the input has infinite variance. The preceding theory enables us to evaluate the mean-square value of the output for both problems. As a matter of convenience, we do this in the real frequency domain rather than the complex s domain.

Problem 1:

$$S_x(j\omega) = \begin{cases} \frac{A}{1 + (T\omega)^2}, & |\omega| \leq \omega_c \\ 0, & |\omega| > \omega_c \end{cases} \quad (3.4.5)$$

$$E(x^2) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} \frac{A}{1 + (T\omega)^2} d\omega = \frac{A}{\pi T} \tan^{-1}(\omega_c T) \quad (3.4.6)$$

Problem 2:

$$S_x(j\omega) = \frac{A}{1 + (T\omega)^2}, \quad \text{for all } \omega \quad (3.4.7)$$

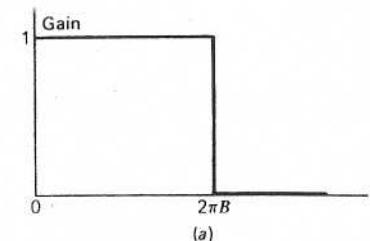
$$\begin{aligned} E(x^2) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{A}{1 + (T\omega)^2} d\omega \\ &= \frac{A}{\pi T} \tan^{-1}(\infty) = \frac{A}{2T} \end{aligned} \quad (3.4.8)$$

Now, we see by comparing the results given by Eqs. (3.4.6) and (3.4.8) that the difference is just that between $\tan^{-1}(\omega_c T)$ and $\tan^{-1}(\infty)$. The bandwidth of the input is ω_c and the filter bandwidth is $1/T$. Thus, if their ratio is large, $\tan^{-1}(\omega_c T) \approx \tan^{-1}(\infty)$. For a ratio of 100:1, the error is less than 1 percent. Thus, if the input spectrum is flat considerably out beyond the point where the system response is decreasing at 20 db/decade (or faster), there is relatively little error introduced by assuming the input is flat out to infinity. The resulting simplification in the analysis is significant.

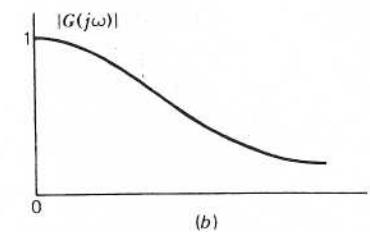
3.5 NOISE EQUIVALENT BANDWIDTH

In filter theory, it is sometimes convenient to think of an idealized filter whose frequency response is unity over a prescribed bandwidth B (in hertz) and zero outside this band. This response is depicted in Fig. 3.3a. If this ideal filter is driven by white noise with amplitude A , its mean-square response is

$$E(x^2)(\text{ideal}) = \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} Ad\omega = 2AB \quad (3.5.1)$$



(a)



(b)

Figure 3.3. Ideal and actual filter responses. (a) Ideal. (b) Actual.

Next, consider an actual filter $G(s)$ whose gain has been normalized to yield a peak response of unity. An example is shown in Fig. 3.3b. The mean-square response of the actual filter to white noise of amplitude A is given by

$$E(x^2)(\text{actual}) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} AG(s)G(-s) ds \quad (3.5.2)$$

Now, if we wish to find the idealized filter that will yield this same response, we simply equate $E(x^2)(\text{ideal})$ and $E(x^2)(\text{actual})$ and solve for the bandwidth that gives equality. The resultant bandwidth B is known as the *noise equivalent bandwidth*. It may, of course, be written explicitly as

$$B(\text{in hertz}) = \frac{1}{2} \left[\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} G(s)G(-s) ds \right] \quad (3.5.3)$$

Example 3.4 Suppose we wish to find the noise equivalent bandwidth of the second-order low-pass filter

$$G(s) = \frac{1}{(1 + Ts)^2}$$

Since the peak response of $G(s)$ occurs at zero frequency and is unity, the gain scale factor is set properly. We must next evaluate the integral in brackets in Eq. (3.5.3). Clearly, $G(s)$ is second-order in the denominator, and therefore we use I_2 of the integral tables given in Section 3.3. The coefficients in this case are

$$c_1 = 0 \quad d_2 = T^2$$

$$c_0 = 1 \quad d_1 = 2T$$

$$d_0 = 1$$

and thus I_2 is

$$I_2 = \frac{c_0^2}{2d_0 d_1} = \frac{1}{4T}$$

The filter's noise equivalent noise bandwidth is then

$$B = \frac{1}{8T} \text{ Hz}$$

This says, in effect, that an idealized filter with a bandwidth of $1/8T$ Hz would pass the same amount of noise as the actual second-order filter. ■

3.6 SHAPING FILTER

With reference to Fig. 3.4, we have seen that the output spectral function can be written as

$$S_x(s) = 1 \cdot G(s)G(-s) \quad (3.6.1)$$

If $G(s)$ is minimum phase and rational in form,* Eq. (3.6.1) immediately provides a factored form for $S_x(s)$ with poles and zeros automatically separated into left and right half-plane parts.

Clearly, we can reverse the analysis problem and pose the question: What minimum-phase transfer function will shape unity white noise into a given spectral function $S_x(s)$? The answer should be apparent. If we can use spectral factorization on $S_x(s)$, the part with poles and zeros in the left half plane provides the appropriate shaping filter. This is a useful concept, both as a mathematical artifice and also as a physical means of obtaining a noise source with desired spectral characteristics from a wideband source.

Example 3.5 Suppose we wish to find the shaping filter that will shape unity white noise into noise with a spectral function

$$S_x(j\omega) = \frac{\omega^2 + 1}{\omega^4 + 64} \quad (3.6.2)$$

First, we write S_x in the s domain as

$$S_x(s) = \frac{-s^2 + 1}{s^4 + 64} \quad (3.6.3)$$

Next, we find the poles and zeros of S_x .

$$\text{Zeros} = \pm 1$$

$$\text{Poles} = -2 \pm j2, \quad 2 \pm j2$$

*This condition requires $G(s)$ to have a finite number of poles and zeros, all of which must be in the left half-plane.

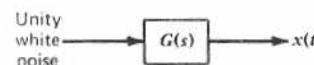


Figure 3.4. Shaping filter.

Finally, we group together left and right half-plane parts. $S_x(s)$ can then be written as

$$S_x(s) = \frac{s + 1}{s^2 + 4s + 8} \cdot \frac{-s + 1}{s^2 - 4s + 8} \quad (3.6.4)$$

The desired shaping filter is then

$$G(s) = \frac{s + 1}{s^2 + 4s + 8} \quad (3.6.5)$$

3.7 NONSTATIONARY (TRANSIENT) ANALYSIS—INITIAL CONDITION RESPONSE

As mentioned previously, the response of a linear system may always be considered as a superposition of an initial-condition part and a driven part. The response due to the initial conditions is often ignored in tutorial discussions, but it should not be because there are many applications where the initial conditions are properly modeled as random variables. If this is the case, one simply solves the problem using standard deterministic methods leaving the initial conditions in the solution in general terms. An example will illustrate the procedure.

Example 3.6 Suppose the circuit shown in Fig. 3.5 has been in operation for a long time with the switch open, and then the switch is closed at a random time which we denote as $t = 0$. We are asked to describe the voltage across the capacitor after $t = 0$.

We first look at the steady-state condition just prior to closing the switch.

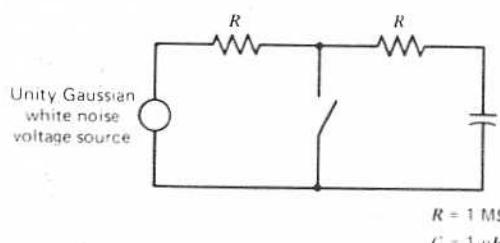


Figure 3.5. Circuit for Example 3.6.

The transfer function relating the capacitor voltage to the white noise source is given by

$$G(s) = \frac{1}{1 + 2s} \quad (3.7.1)$$

Therefore, using the methods of Sections 3.2 and 3.3, the mean-square value of the capacitor voltage v_c is found to be

$$E(v_c^2) = \frac{1}{2\pi j} \int_{-\infty}^{j\infty} \frac{1}{1 + 2s} \cdot \frac{1}{1 - 2s} ds = \frac{1}{4} \quad (3.7.2)$$

We have now established that the initial condition is a normal random variable with zero mean and a variance of $\frac{1}{4}$.

After the switch is closed the system differential equation is

$$C\dot{v}_c + \frac{1}{R} v_c = 0 \quad (3.7.3)$$

Taking the Laplace transform of both sides yields

$$C[sV_c(s) - v_c(0)] + \frac{1}{R} V_c(s) = 0$$

or

$$V_c(s) = \frac{v_c(0)}{s + 1/RC} \quad (3.7.4)$$

The explicit expression for the time-domain waveform is now obtained by taking the inverse transform of Eq. (3.7.4). It is

$$v_c(t) = v_c(0)e^{-t/RC} \quad (3.7.5)$$

where $v_c(0)$ is a random variable characterized by $N(0, \frac{1}{4})$. Note the solution of the initial-condition problem leads to a deterministic random process; that is, the process has deterministic structure and any particular realization of the process is exactly predictable once the initial condition is known.

The mean-square value of a random process is usually of prime interest. In this case it is easily computed as

$$\begin{aligned} E(v_c^2) &= E[v_c(0)e^{-t/RC}]^2 \\ &= e^{-2t/RC} \cdot E[v_c^2(0)] \\ &= \frac{1}{4} e^{-2t/RC} \end{aligned} \quad (3.7.6)$$

It is, of course, a function of time. ■

The extension of the procedure of Example 3.6 to more complicated situations is fairly obvious, so this will not be pursued further.

3.8 NONSTATIONARY (TRANSIENT) ANALYSIS—FORCED RESPONSE

The block diagram of Fig. 3.1 is repeated as Fig. 3.6 with the addition of a switch in the input. Imagine the system to be initially at rest, and then close the switch at $t = 0$. A transient response takes place in the stochastic problem just as in the corresponding deterministic problem. If the input $f(t)$ is a nondeterministic random process, we would expect the response also to be nondeterministic, and its autocorrelation function may be computed in terms of the input autocorrelation function. This is done as follows.

The system response can be written as a convolution integral

$$x(t) = \int_0^t g(u)f(t-u) du \quad (3.8.1)$$

where $g(u)$ is the inverse Laplace transform of $G(s)$ and is usually referred to as the system weighting function. To find the autocorrelation function we simply evaluate $E[x(t_1)x(t_2)]$.

$$\begin{aligned} R_x(t_1, t_2) &= E[x(t_1)x(t_2)] \\ &= E \left[\int_0^{t_1} g(u)f(t_1-u) du \int_0^{t_2} g(v)f(t_2-v) dv \right] \\ &= \int_0^{t_2} \int_0^{t_1} g(u)g(v) E[f(t_1-u)f(t_2-v)] du dv \end{aligned} \quad (3.8.2)$$

Now, if $f(t)$ is stationary, Eq. (3.8.2) can be written as

$$R_x(t_1, t_2) = \int_0^{t_2} \int_0^{t_1} g(u)g(v)R_f(u-v+t_2-t_1) du dv \quad (3.8.3)$$

and we now have an expression for the output autocorrelation function in terms of the input autocorrelation function and system weighting function.

Equation (3.8.3) is difficult to evaluate except for relatively simple systems. Thus, we are often willing to settle for less information about the response and just compute its mean-square value. This is done by letting $t_2 = t_1 = t$ in Eq. (3.8.3) with the result

$$E[x^2(t)] = \int_0^t \int_0^t g(u)g(v)R_f(u-v) du dv \quad (3.8.4)$$

Three examples will now illustrate the use of Eqs. (3.8.3) and (3.8.4).

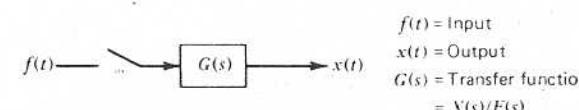


Figure 3.6. Block diagram for nonstationary analysis problem.

Example 3.7 Let $G(s)$ be a first-order low-pass filter, and let $f(t)$ be white noise with amplitude A . Then

$$G(s) = \frac{1}{1 + Ts}$$

$$S_f(\omega) = A$$

Taking inverse transforms gives

$$g(u) = \frac{1}{T} e^{-u/T}$$

$$R_f(\tau) = A \delta(\tau)$$

Next, substituting in Eq. (3.8.4) yields

$$\begin{aligned} E[x^2(t)] &= \int_0^t \int_0^t \frac{A}{T^2} e^{-u/T} e^{-v/T} \delta(u - v) du dv \\ &= \frac{A}{T^2} \int_0^t e^{-2v/T} dv \\ &= \frac{A}{2T} [1 - e^{-2t/T}] \end{aligned} \quad (3.8.5)$$

Note that as $t \rightarrow \infty$, the mean-square value approaches $A/2T$ which is the same result obtained in Section 3.2 using spectral analysis methods. ■

Example 3.8 Let $G(s)$ be an integrator with zero initial conditions, and let $f(t)$ be a Gauss-Markov process with variance σ^2 and time constant $1/\beta$. We desire the mean-square value of the output x . The transfer function and input autocorrelation function are

$$G(s) = \frac{1}{s} \quad \text{or} \quad g(u) = 1$$

and

$$R_f(\tau) = \sigma^2 e^{-\beta|\tau|}$$

Next, we use Eq. (3.8.4) to obtain $E[x^2(t)]$.

$$E[x^2(t)] = \int_0^t \int_0^t 1 \cdot 1 \cdot \sigma^2 e^{-\beta|u-v|} du dv \quad (3.8.6)$$

Some care is required in evaluating Eq. (3.8.6) because one functional expression for $e^{-\beta|u-v|}$ applies for $u > v$, and a different one applies for $u < v$. This is shown in Fig. 3.7. Recognizing that the region of integration must be split into two parts, we have

$$E[x^2(t)] = \int_0^t \int_0^v \sigma^2 e^{-\beta(v-u)} du dv + \int_0^t \int_v^t \sigma^2 e^{-\beta(u-v)} du dv \quad (3.8.7)$$

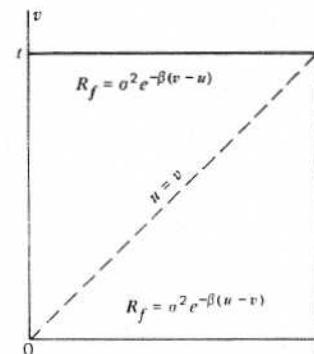


Figure 3.7. Regions of integration for Example 3.8.

Since there is symmetry in the two integrals of Eq. (3.8.7), we can simply evaluate the first one and multiply by 2. The mean-square value of x is then

$$\begin{aligned} E[x^2(t)] &= 2 \int_0^t \int_0^v \sigma^2 e^{-\beta(v-u)} du dv = 2 \int_0^t \sigma^2 e^{-\beta v} \int_0^v e^{\beta u} du dv \\ &= \frac{2\sigma^2}{\beta} \int_0^t e^{-\beta v} (e^{\beta v} - 1) dv \\ &= \frac{2\sigma^2}{\beta^2} [\beta t - (1 - e^{-\beta t})] \end{aligned} \quad (3.8.8)$$

Note that $E[x^2(t)]$ increases without bound as $t \rightarrow \infty$. This might be expected because an integrator is an unstable system. ■

Example 3.9 As our final example, we find the autocorrelation function of the output of a simple integrator driven by unity-amplitude Gaussian white noise. The transfer function and input autocorrelation function are

$$G(s) = \frac{1}{s} \quad \text{or} \quad g(u) = 1$$

$$R_f(\tau) = \delta(\tau)$$

We obtain $R_x(t_1, t_2)$ from Eq. (3.8.3):

$$R_x(t_1, t_2) = \int_0^{t_2} \int_{t_1}^{t_2} 1 \cdot 1 \cdot \delta(u - v + t_2 - t_1) du dv \quad (3.8.9)$$

The region of integration for this double integral is shown in Fig. 3.8 for $t_2 > t_1$.

The argument of the Dirac delta function in Eq. (3.8.9) is zero along the dashed line in the figure. It can be seen that it is convenient to integrate first with respect to v if $t_2 > t_1$ as shown in the figure. Considering this case first (i.e., $t_2 > t_1$), we have

$$R_x(t_1, t_2) = \int_0^{t_1} \int_{t_1}^{t_2} \delta(u - v + t_2 - t_1) dv du = \int_0^{t_1} 1 \cdot du = t_1$$

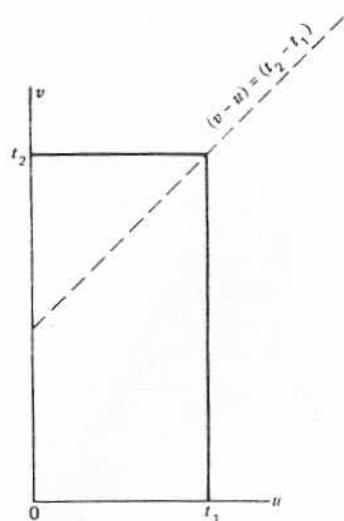


Figure 3.8. Region of integration for Example 3.9.

Similarly, when $t_2 < t_1$,

$$R_x(t_1, t_2) = t_2$$

The final result is then

$$R_x(t_1, t_2) = \begin{cases} t_1, & t_2 \geq t_1 \\ t_2, & t_2 < t_1 \end{cases} \quad (3.8.10)$$

Note that this is the same result obtained for the Wiener process in Chapter 2. ■

In concluding this section, we might comment that if the transient response includes both forced and initial-condition components, the total response is just the superposition of the two. The mean-square value must be evaluated with care, though, because the total mean-square value is the sum of the two only when the crosscorrelation is zero. If the crosscorrelation between the two responses is not zero, it must be properly accounted for in computing the mean-square value.

3.9 SUMMARY

The mean-square value and spectral density function (or autocorrelation function) are the output process descriptors of prime interest. This is partially a matter of mathematical convenience, because they are usually the only descriptors that can be readily computed. If only the stationary solution is desired, the

output spectral density is readily computed from Eq. (3.2.1) or (3.2.2). The mean-square value may then be obtained from the integral of the spectral function. Integral tables are available to assist in this task, provided the transfer and input spectral functions are both in rational form.

In transient problems, the portion of the system response due to random initial conditions is computed using standard deterministic methods. The resulting response is always a deterministic random process. The autocorrelation function of the driven response may be computed from Eq. (3.8.3). In many cases the computation is quite involved; therefore, only the mean-square value is computed. This computation is relatively simple and is given by Eq. (3.8.4). The total response in the transient problem is, of course, the superposition of both the initial-condition and driven parts.

Linear multiple-input, multiple-output problems were not discussed in this chapter. Complicated problems of this type are best handled by state-variable methods, and discussion of such systems will be deferred to Chapter 5. Simple multiple-input, multiple-output problems are sometimes manageable using scalar methods, and in these cases one simply uses superposition in computing the various responses. One must, of course, be careful in computing spectral functions and mean-square values to account properly for any nontrivial crosscorrelations that may exist.

Problems

3.1. In Section 3.1 the equation relating the input and output spectral densities [$S_x(j\omega) = |G(j\omega)|^2 S_f(j\omega)$] was justified with heuristic arguments. This can be formalized by proceeding through the following steps:

- Write the output $x(t)$ as a convolution integral using Fourier rather than Laplace transforms.
- Do likewise for the shifted output $x(t + \tau)$.
- Multiply the expressions for $x(t)$ and $x(t + \tau)$ and (symbolically) form the expectation of the product.
- Now note that the Fourier transform of the autocorrelation function is the spectral density; transform both sides, interchange the order of integration and the desired result is apparent.

Proceed through the steps just described and formally justify the $S_x(j\omega) = |G(j\omega)|^2 S_f(j\omega)$.

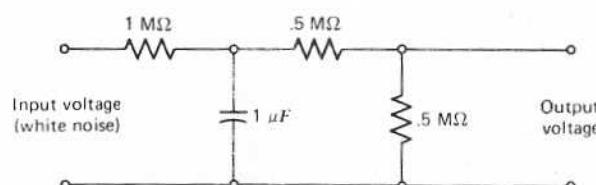
3.2. Find the steady-state mean-square value of the output for the following filters. The input is white noise with a spectral density amplitude A .

$$(a) G(s) = \frac{Ts}{(1 + Ts)^2}$$

$$(b) G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

$$(c) G(s) = \frac{s + 1}{(s + 2)^2}$$

- 3.3.** A white-noise process having a spectral density amplitude of A is applied to the circuit shown. The circuit has been in operation for a long time. Find the mean square value of the output voltage.

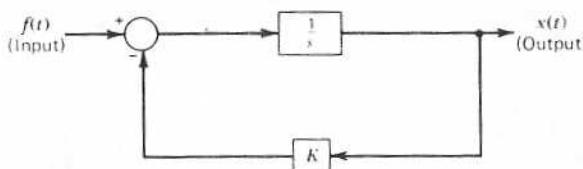
**Problem 3.3**

- 3.4.** The input to the feedback system shown is a stationary Markov process with an autocorrelation function

$$R_f(\tau) = \sigma^2 e^{-\beta|\tau|}$$

The system is in a stationary condition.

- (a) What is the spectral density function of the output?
- (b) What is the mean-square value of the output?

**Problem 3.4**

- 3.5.** Consider the nonminimum phase filter

$$G(s) = \frac{1 - T_1 s}{1 + T_2 s}$$

driven with a stationary Gauss-Markov process with an autocorrelation function $R(\tau) = \sigma^2 e^{-\beta|\tau|}$. Find:

- (a) The spectral density function of the output.
- (b) The mean-square value of the output.

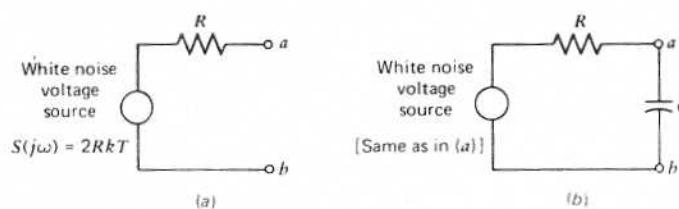
- 3.6.** Find the steady-state mean-square value of the output for a first-order low-pass filter [i.e., $G(s) = 1/(1 + Ts)$] if the input has an autocorrelation function of the form

$$R(\tau) = \begin{cases} \sigma^2(1 - \beta|\tau|), & -\frac{1}{\beta} \leq \tau \leq \frac{1}{\beta} \\ 0, & |\tau| > 1/\beta \end{cases}$$

(Hint: The input spectral function is irrational so the integrals given in Table 3.1 are of no help here. One approach is to write the integral expression for $E(X^2)$ in terms of real ω rather than s and then use conventional integral tables. Also, those familiar with residue theory will find that the integral can be evaluated by the method of residues.)

- 3.7.** Thermal noise in a metallic resistor is sometimes modeled as a white noise voltage source in series with the resistance R of the resistor (7.8). This is shown in the accompanying figure, along with the parameters describing the spectral amplitude of the noise source. At room temperature the flat spectrum approximation is reasonably accurate from zero frequency out to the infrared range. Clearly, in the idealized model of part (a), the voltage from a to b would be infinity, which is physically impossible. The model is still useful, though, because there is always some shunt capacitance associated with the load connected from a to b . If nothing else, the parasitic capacitance of the resistor leads is sufficient to cause the spectral function to "roll off" at 20 db/decade and thus cause the output to be bounded. This is shown in part (b) of the figure. Now, to demonstrate the validity of the white noise model, consider an example where $R = 1 \times 10^6 \Omega$ and $C = 1 \times 10^{-12} F$ (a plausible value for parasitic capacitance). Also, assume the temperature is room temperature, about 290 K.

- (a) Find the rms voltage across the capacitor C .
- (b) Find the half-power frequency in hertz. (Defined as the frequency at which the spectral function is half its value at zero frequency.)
- (c) Based on the result of part (b), would you think it reasonable to consider this noise source as being flat (i.e., pure white) in most electronic-circuit applications? Explain briefly.



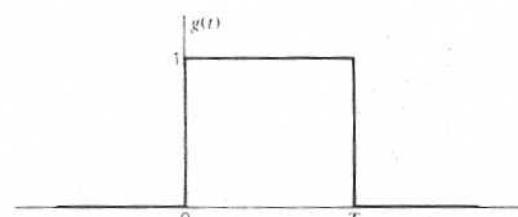
$$R = \text{Resistance of metallic resistor (ohms)}$$

$$k = \text{Boltzmann constant } \approx 1.38 \times 10^{-23} \text{ (joules/deg)}$$

$$T = \text{Temperature (degrees Kelvin)}$$

Problem 3.7

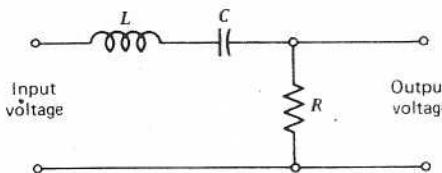
- 3.8.** Consider a linear filter whose weighting function is shown in the figure. (This filter is sometimes referred to as a finite-time integrator.) The input to the filter is white noise with a spectral density amplitude A , and the filter has been in operation a long time. What is the mean-square value of the output?

**Problem 3.8.** Filter weighting function

3.9. Find the shaping filter that will shape unity white noise into noise with a spectral function

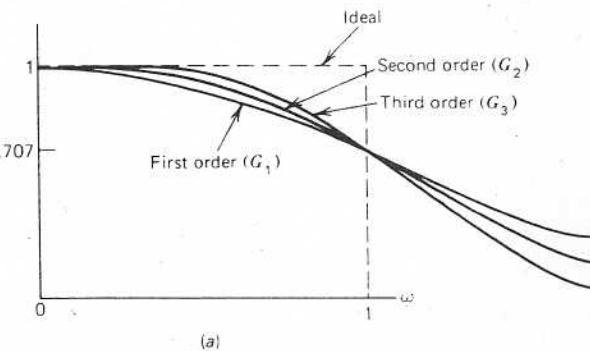
$$S(j\omega) = \frac{\omega^2 + 1}{\omega^4 + 8\omega^2 + 16}$$

3.10. A series resonant circuit is shown in the figure. Let the resistance R be small such that the circuit is sharply tuned (i.e., high Q or very low damping ratio). Find the noise equivalent bandwidth for this circuit and express it in terms of the damping ratio ζ and the natural undamped resonant frequency ω_r (i.e., $\omega_r = 1/\sqrt{LC}$). Note that the "ideal" response in this case is a unity-gain rectangular pass band centered about ω_r . Also find the usual half-power bandwidth and compare this with the noise equivalent bandwidth. (Half-power bandwidth is defined to be the frequency difference between the two points on the response curve that are "down" by a factor of $1/\sqrt{2}$ from the peak value. It is useful to approximate the resonance curve as being symmetric about the peak for this part of the problem.)



Problem 3.10

3.11. The transfer functions and corresponding bandpass characteristics for first-, second-, and third-order Butterworth filters are shown in the accompanying figure (9).



(a)

$$G_1(s) = \frac{1}{s+1}$$

$$G_2(s) = \frac{1}{s^2 + \sqrt{2}s + 1}$$

$$G_3(s) = \frac{1}{(s+1)(s^2 + s + 1)}$$

(b)

Problem 3.11. (a) Responses of three Butterworth filters. (b) Transfer functions of Butterworth filters.

These filters are said to be "maximally flat" at zero frequency with each successive higher-order filter more nearly approaching the ideal curve than the previous one. All three filters have been normalized such that all responses intersect the -3-db point at 1 rad/sec (or $1/2\pi$ Hz).

- (a) Find the noise equivalent bandwidth for each of the filters.
 (b) Insofar as noise suppression is concerned, is there much to be gained by using anything higher order than a third-order Butterworth filter?

3.12. Find the mean-square value of the output (averaged in an ensemble sense) for the following transfer functions. In both cases, the initial conditions are zero and the input $f(t)$ is applied at $t = 0$.

$$(a) G(s) = \frac{1}{s^2}, \quad R_f(\tau) = A \delta(\tau)$$

$$(b) G(s) = \frac{1}{s^2 + \omega_0^2}, \quad R_f(\tau) = A \delta(\tau)$$

3.13. A certain linear system is known to satisfy the following differential equation:

$$\ddot{x} + a\dot{x} = f(t)$$

$$x(0) = \dot{x}(0) = 0$$

where $x(t)$ is the response, and $f(t)$ is the input that is applied at $t = 0$. If $f(t)$ is white noise with spectral density amplitude A , what is the mean-square value of the response $x(t)$?

3.14. Consider a simple first-order low-pass filter whose transfer function is

$$G(s) = \frac{1}{1 + 10s}$$

The input to the filter is initiated at $t = 0$, and the filter's initial condition is zero. The input is given by

$$f(t) = A u(t) + n(t)$$

where

$u(t)$ = unit step function

A = random variable with uniform distribution from 0 to 1

$n(t)$ = unity Gaussian white noise

Find:

- (a) The mean, mean square, and variance of the output evaluated at $t = .1$ sec.
 (b) Repeat (a) for the steady-state condition (i.e., for $t = \infty$).

(Hint: Since the system is linear, superposition may be used in computing the output. Note that the deterministic component of the input is written explicitly in functional form. Therefore, deterministic methods may be used to compute the portion of the output due to this component. Also, remember that the mean square value and the variance are not the same if the mean is nonzero.)

3.15. A signal is known to have the following form:

$$s(t) = a_0 + n(t)$$

where a_0 is an unknown constant, and $n(t)$ is a stationary noise process with a known autocorrelation function

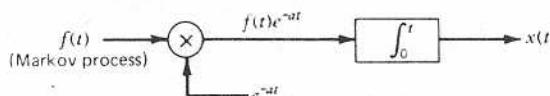
$$R_n(\tau) = \sigma^2 e^{-\beta|\tau|}$$

It is suggested that a_0 can be estimated by simply averaging $s(t)$ over a finite interval of time T . What would be the rms error in the determination of a_0 by this method? (Note that root mean square rather than mean square value is requested in this problem.)

3.16. In the figure shown, $f(t)$ is a time stationary random process whose autocorrelation function is

$$R_f(\tau) = \sigma^2 e^{-\beta|\tau|}$$

The input $f(t)$ is first multiplied by e^{-at} and then integrated, beginning at $t = 0$. The initial value of the integrator is zero. Find the mean-square value of $x(t)$.

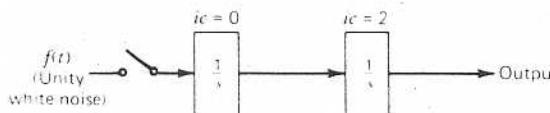


Problem 3.16

3.17. Consider an integrator whose initial output at $t = 0$ is a Gaussian random variable with zero mean and variance σ^2 . A Gaussian white-noise input with spectral amplitude A is applied at $t = 0$. What is the mean-square value of the output as a function of time?

3.18. Unity Gaussian white noise $f(t)$ is applied to the cascaded combination of integrators shown in the figure. The switch is closed at $t = 0$. The initial condition for the first integrator is zero, and the second integrator has two units as its initial value.

- (a) What is the mean square value of the output at $t = 2$ sec?
- (b) Sketch the probability density function for the output evaluated at $t = 2$ sec.



Problem 3.18

3.19. Consider the random process defined by the transfer function shown in the figure. The input $f(t)$ is Gaussian white noise with unity spectral amplitude, and the process is started at $t = 0$ with zero initial conditions. The autocorrelation function of the output $x(t)$ is defined as

$$R_x(t_1, t_2) = E[x(t_1)x(t_2)], \quad t_1 \text{ and } t_2 > 0$$

Find $R_x(t_1, t_2)$.



Problem 3.19

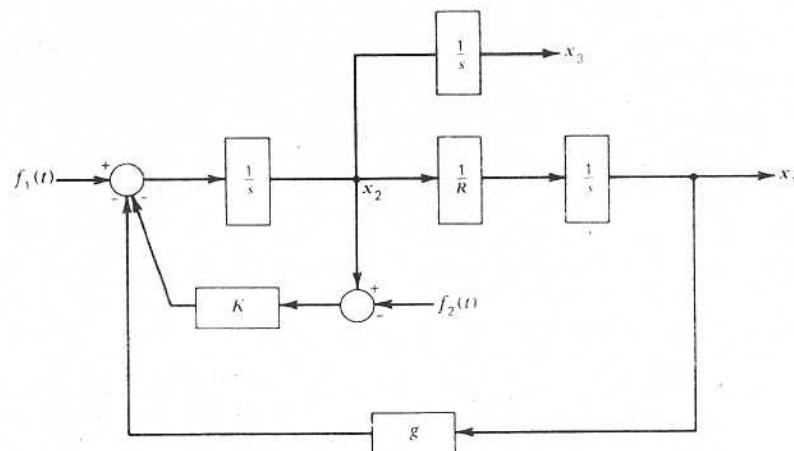
3.20. Consider again the filter with a rectangular weighting function discussed in Problem 3.8. Consider the filter to be driven with unity Gaussian white noise, which is initiated at $t = 0$ with zero initial conditions.

(a) Find the mean-square response in the interval from 0 to T .

(b) Find the mean-square response for $t \geq T$ and compare the result with that obtained in Prob. 3.8.

(c) From the result of (b), would you say the filter's "memory" is finite or infinite?

3.21. The block diagram shown describes the error propagation in one channel of an inertial navigation system with external-velocity-reference damping (10). The inputs shown as $f_1(t)$ and $f_2(t)$ are random driving functions due to the accelerometer and external-velocity-reference instrument errors. These will be assumed to be independent white noise processes with spectral amplitudes A_1 and A_2 , respectively. The outputs are labeled x_1 , x_2 , and x_3 , and these physically represent the inertial system's platform tilt, velocity error, and position error. Find the steady-state mean-square value of each of the outputs. (Hint: Since the system is linear, use superposition and note that the driving functions are independent.)



$R = \text{earth radius} \approx 2.09 \times 10^7 \text{ ft}$

$g = \text{gravitational constant} \approx 32.2 \text{ ft/sec}^2$

$K = \text{feedback constant (adjustable to yield the desired damping ratio)}$

Problem 3.21

3.22. A hypothetical communication channel is shown in the figure. The channel has inherent low-pass characteristics as indicated by $G(s)$. The signal is noiselike in character with a spectral density given by

$$S_f(j\omega) = \begin{cases} 1, & -1 \leq \omega \leq 1 \\ 0, & |\omega| > 1 \end{cases}$$

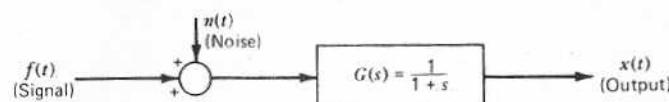
The additive noise is white with a spectral density

$$S_n(j\omega) = .1, \quad -\infty < \omega < \infty$$

It is suggested that the signal distortion can be reduced by inserting the following equalizer (compensator) in cascade with $G(s)$:

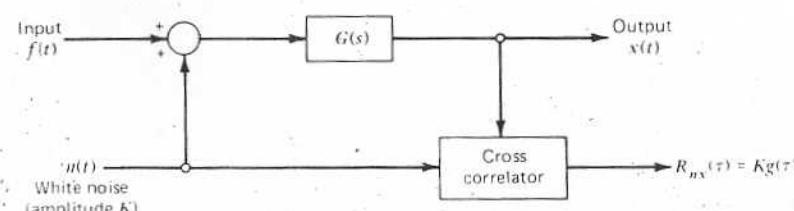
$$G_{eq}(s) = \frac{1+s}{1+.5s}$$

- (a) Considering only the signal component of the output, is the above suggestion valid?
- (b) Define the output signal and noise "powers" as their respective mean-square values. Find the output power signal-to-noise ratio before and after insertion of the equalizer.
- (c) On the basis of (a) and (b), would you say there will be a net improvement in the fidelity of the output waveform by the insertion of the suggested equalizer?



Problem 3.22

3.23. The block diagram in the accompanying figure shows a means of determining the weighting function of a linear system (and thus, indirectly, the system's transfer function). The basic idea is to superimpose a small amount of white noise on the regular input and then crosscorrelate the resulting output with the intentionally added white noise. If the amplitude of the additive noise is relatively small, it is scarcely noticeable, if at all, and this provides a means of continuously monitoring the transfer characteristics of the system without disturbing its normal operation. Show that the output of the crosscorrelator $R_{nx}(\tau)$ is, in fact, proportional to the system weighting function $g(\tau)$. [If you need help on this one, see Truxal (11), p. 437. This idea has "been around" for a long time but has seen only limited application. The reason, of course, lies in the computational effort in determining the crosscorrelation $R_{nx}(\tau)$. This is no small task, especially in on-line applications. Perhaps as the computer state-of-the-art advances, this technique will find more application.]

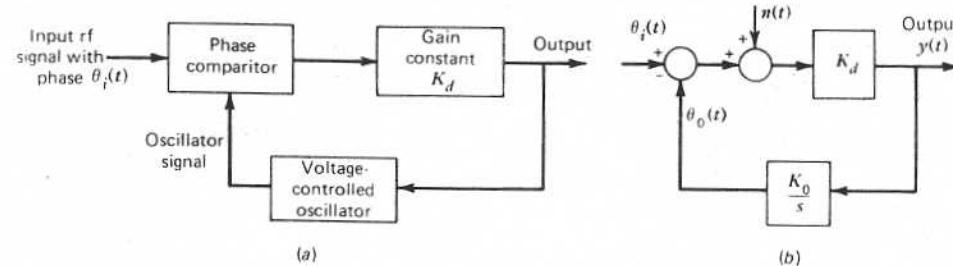


Problem 3.23

3.24. Part (a) of the accompanying figure shows a functional block diagram of an elementary phase-lock loop. Such circuits are used in a wide variety of communication applications (8.12). Briefly, the phase comparator provides a signal proportional to the difference between the phase of the incoming rf signal and that of the local oscillator. The phase difference is modified by gain factor K_d and is then fed back as a voltage to the local voltage-controlled oscillator (VCO). This voltage causes the frequency of the VCO to shift up or down as necessary to lock onto the phase of the incoming rf signal.

Part (b) of the figure shows the linearized block diagram for a phase-lock loop with the addition of a phase noise input labeled $n(t)$. If the incoming rf signal is frequency modulated (FM), it is the derivative of $\theta_i(t)$ that is proportional to the baseband signal, and thus $\dot{\theta}_i(t)$ is the signal to be recovered (detected) in this case.

- (a) Show that the transfer function relating $\dot{\theta}_i(t)$ to the output $y(t)$ is the equivalent of a differentiator in cascade with a first-order low-pass filter.
- (b) Consider the phase noise $n(t)$ to be flat with a spectral amplitude N_0 . Also assume the gain parameters of the phase-lock loop are chosen such that the baseband signal is passed with no distortion, that is, the loop response is flat over the signal frequency range, say from 0 to W Hz. Find the spectral function of the output noise in the 0 to W Hz range, and find its average power (i.e., mean-square value) in this frequency range. (It may help here to think of the noise as being limited to the 0 to W Hz range by a sharp-cutoff postdetector filter.)



Problem 3.24

3.25. For two stationary random processes $x(t)$ and $y(t)$, the coherence function is defined to be (see Chapter 2):

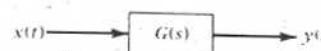
$$\gamma_{xy}^2(\omega) = \frac{|S_{xy}(j\omega)|^2}{S_x(j\omega)S_y(j\omega)} \quad (\text{P3.25.1})$$

Consider a special case where $x(t)$ is the input to a linear system, and $y(t)$ is the output as shown in the accompanying figure.

- (a) First show that

$$S_{xy}(j\omega) = G(j\omega)S_x(j\omega) \quad (\text{P3.25.2})$$

- (b) Then show that the coherence function is unity for all ω . (This shows that x and y do not have to be equal in order to have unity coherence. They only need to be intimately related, as, of course, they are in this situation.)



Problem 3.25

References Cited in Chapter 3

1. J. J. D'Azzo and C. H. Houpis, *Linear Control System Analysis and Design*, 2nd ed., New York: McGraw-Hill, 1981.

2. R. C. Dorf, *Modern Control Systems*, 3rd ed., Reading, Mass.: Addison-Wesley, 1980.
3. R. A. Gabel and R. A. Roberts, *Signals and Linear Systems*, 2nd ed., New York: Wiley, 1980.
4. H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of Servomechanisms*, Radiation Laboratory Series (Vol. 25), New York: McGraw-Hill, 1947.
5. G. C. Newton, L. A. Gould, and J. F. Kaiser, *Analytical Design of Linear Feedback Controls*, New York: Wiley, 1957.
6. G. R. Cooper and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, New York: Holt, Rinehart, and Winston, 1971.
7. A. B. Carlson, *Communication Systems*, 2nd ed., New York: McGraw-Hill, 1975.
8. K. S. Shanmugam, *Digital and Analog Communication Systems*, New York: Wiley, 1979.
9. D. Childers and A. Durling, *Digital Filtering and Signal Processing*, St. Paul: West Publishing, 1975.
10. G. R. Pitman (Ed.), *Inertial Guidance*, New York: Wiley, 1962.
11. J. G. Truxal, *Automatic Feedback Control System Synthesis*, New York: McGraw-Hill, 1955.
12. F. M. Gardner, *Phaselock Techniques*, 2nd ed., New York: Wiley, 1979.

Additional References for General Reading

13. P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles*, New York: McGraw-Hill, 1980.
14. H. J. Larson and B. O. Shubert, *Probabilistic Models in Engineering Sciences* (Vols. 1 and 2), New York: Wiley, 1979.

CHAPTER 4

Wiener Filtering

In this and subsequent chapters we will consider a particular branch of filter theory that is sometimes referred to as least-squares filtering. Actually, this is an oversimplification because it is the *average* squared error that is minimized and not just the squared error. "Linear minimum mean-square error filtering" would be a more descriptive name for this type of filtering. This is a bit wordy, though, so the name is usually shortened and the theory is included as part of the general theory of least squares. Simply stated, the least-squares filter problem is this: Given the spectral characteristics of an additive combination of signal and noise, what linear operation on this input combination will yield the best separation of the signal from the noise? "Best" in this case means minimum mean-square error. This branch of filtering began with N. Wiener's work in the 1940s (1). R. E. Kalman then made an important contribution in the early 1960s by providing an alternative approach to the same problem using state space methods (2,3). Kalman's contribution has been especially significant in applied work, because his solution is readily implemented with modern digital methods.

We will consider the Wiener and Kalman theories in their historical order. It should be mentioned that neither is prerequisite material for the other; therefore, they may be studied in either order, or one to the exclusion of the other, for that matter.

4.1 PERSPECTIVE

The purpose of any filter is to separate one thing from another. In the electric filter case, this usually refers to passing signals in a specified frequency range and rejecting those outside that range; and, historically, filter theory began with the problem of designing a circuit to yield the desired frequency response. This is still an important problem. In many applications in communication and control, one knows intuitively what the ideal frequency response should be. For example, if we want to receive the signal from a particular AM radio station (and do it faithfully), we know that the appropriate filter is one that passes all frequencies within a few kilohertz on either side of the assigned station frequency and rejects all others. Certainly no elaborate theory is needed to determine the desired frequency response in this case. The problem is simply one of circuit design. We will see, though, that this is not always the case. During

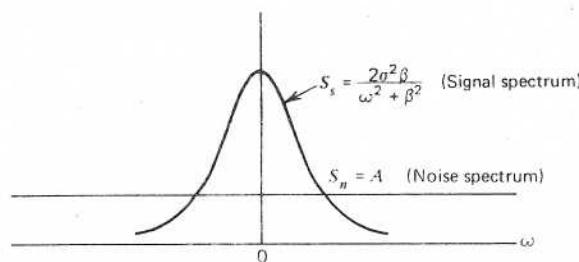


Figure 4.1. Spectral densities of signal and noise.

World War II, Norbert Wiener considered a different sort of filter problem (1). Suppose the signal, as well as the noise, is noiselike in character, and suppose further that there is a significant overlap in the spectra of both the signal and noise. For example, say the signal is a Gauss-Markov process and the corrupting noise is white noise. Their spectral densities are shown in Fig. 4.1. Now, in this case, it should be apparent that no filter is going to yield perfect separation, and the filter that gives the best compromise of passing the signal and, at the same time, suppresses most of the noise is not at all obvious. Neither is it obvious how one should define "best compromise" in order to make the problem mathematically tractable. This is the problem Wiener examined in the 1940s. We note that he was not concerned with filter design in the sense of choosing appropriate resistors, capacitors, and so forth. Instead, his problem was more fundamental: namely, what should the filter's frequency response be in order to give the best possible separation of signal from noise?

The theory that is now loosely referred to as Wiener filter theory is characterized by:

1. The assumption that both signal and noise are random processes with known spectral characteristics or, equivalently, known auto- and cross-correlation functions.
2. The criterion for best performance is minimum mean-square error. This is partially to make the problem mathematically tractable, but it is also a good physical criterion in many applications.
3. A solution based on scalar methods that leads to the optimal filter weighting function (or transfer function in the stationary case).

We now proceed to the filter optimization problem.

4.2 OPTIMIZATION WITH RESPECT TO A PARAMETER

One of the early methods of stochastic filter optimization was presented by R. S. Phillips in Volume 25 of the now-famous Radiation Laboratory Series (4).

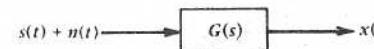


Figure 4.2. Filter optimization problem.

His approach was less general than Wiener's but it is still useful in many applications. Basically, the method is to choose the form of the filter transfer function intuitively, leaving one or more parameters free to vary. One then minimizes the mean-square error with respect to these parameters.

Before looking at an example of Phillips' procedure, we need to derive an expression for the filter's mean-square error. In terms of its Laplace transform, the output of the filter shown in Fig. 4.2. is*

$$X(s) = G(s)[S(s) + N(s)] \quad (4.2.1)$$

We define the filter error as the difference between the actual output and what we would like it to be ideally—the signal. Therefore let†

$$e(t) = s(t) - x(t) \quad (4.2.2)$$

and

$$E(s) = S(s) - X(s) \quad (4.2.3)$$

Substituting Eq. (4.2.1) into Eq. (4.2.3) yields

$$E(s) = S(s) - G(s)[S(s) + N(s)] = [1 - G(s)]S(s) - [G(s)]N(s) \quad (4.2.4)$$

It can be seen that the error can be thought of as a superposition of two components; one due to the signal modified by the transfer function $[1 - G(s)]$ and another due to the noise modified by $-G(s)$. If the signal and noise are uncorrelated, the mean-square error is obtained as simply the sum of two terms, that is,

$$\begin{aligned} E(e^2) &= \frac{1}{2\pi j} \int_{-\infty}^{\infty} [1 - G(s)][1 - G(-s)]S(s) ds \\ &\quad + \frac{1}{2\pi j} \int_{-\infty}^{\infty} G(s)G(-s)N(s) ds \end{aligned} \quad (4.2.5)$$

We now have an explicit expression for the mean-square error in terms of the spectral functions of $s(t)$ and $n(t)$ (presumably known) and the filter transfer function. If $G(s)$ contains a parameter free to vary, we can now use ordinary

*Where there might be confusion between the signal variable $s(t)$ and the complex variable s , we will show the time dependence explicitly, that is, $s(t)$ is the signal variable. Also, uppercase $S(s)$ without a subscript will denote the Laplace transform of $s(t)$; S with a subscript, for example, $S_s(s)$, denotes spectral density of the subscript variable. This overlap in notation should not be confusing when taken within the context of the subject under consideration.

†Some authors prefer to define the error with the opposite sign. In the subsequent optimization we will always be concerned with minimizing the mean-square error. Thus the sign of the error is of no consequence; the resulting optimal filter and mean-square error is the same either way.

differential calculus to minimize $E(e^2)$ with respect to the parameter. We now proceed with an example.

Example 4.1 Consider the Gauss-Markov signal and white noise situation shown in Fig. 4.1. It is apparent that some sort of low-pass filter is needed to separate signal from noise. Let us try a simple first-order filter of the form

$$G(s) = \frac{1}{1 + Ts}$$

We have now specified the functional form for $G(s)$, and hence we are ready to use Eq. (4.2.5). The needed quantities are

$$\begin{aligned} G(s) &= \frac{1}{1 + Ts} = \frac{1/T}{s + 1/T} \\ 1 - G(s) &= 1 - \frac{1}{1 + Ts} = \frac{s}{s + 1/T} \\ S_s(s) &= \frac{2\sigma^2\beta}{-s^2 + \beta^2} = \frac{\sqrt{2\sigma^2\beta}}{s + \beta} \cdot \frac{\sqrt{2\sigma^2\beta}}{-s + \beta} \\ S_n(s) &= A = \sqrt{A} \cdot \sqrt{A} \end{aligned}$$

Substituting the above quantities into Eq. (4.2.5) and evaluating $E(e^2)$ using the integral tables of Section 3.4 yields

$$E(e^2) = \frac{\sigma^2\beta T}{1 + \beta T} + \frac{A}{2T} \quad (4.2.6)$$

This can now be minimized with respect to T using differential calculus. The result is that $E(e^2)$ is a minimum for

$$T = \frac{\sqrt{A}}{\sigma\sqrt{2\beta} - \beta\sqrt{A}} \quad (4.2.7)$$

It is interesting to note that this will yield a positive value of T only for certain values of the parameters. A negative solution for T simply means that no relative minimum exists within the interval from zero to infinity.

It should be remembered that the minimum obtained is not the absolute minimum possible (unless by coincidence), because the form of the filter transfer function was chosen intuitively. Other functional forms might have done better. ■

4.3 THE STATIONARY OPTIMIZATION PROBLEM—WEIGHTING FUNCTION APPROACH*

We now consider the filter optimization problem that Wiener first solved in the 1940s (1). Referring to Fig. 4.3, we assume the following:

1. The filter input is an additive combination of signal and noise, both of which are covariance stationary with known auto- and cross-correlation functions (or corresponding spectral functions).
2. The filter is linear and not time varying. No further assumption is made as to its form.
3. The output is covariance stationary. (A long time has elapsed since any switching operation.)
4. The performance criterion is minimum mean-square error, where the error is defined as $e(t) = s(t + \alpha) - x(t)$.

In addition to the generalization relative to the form of the filter transfer function, we are also generalizing by saying the ideal filter output is to be $s(t + \alpha)$ rather than just $s(t)$. The following terminology has evolved relative to the choice of the α parameter:

1. α positive: This is called the *prediction* problem. (The filter is trying to predict the signal value α units ahead of the present time t .)
2. $\alpha = 0$: This is called the *filter* problem. (The usual problem we have considered before.)
3. α negative: This is called the *smoothing* problem. (The filter is trying to estimate the signal value α units in the past.)

This is an important generalization and there are numerous physical applications corresponding to all three cases. The α parameter is chosen to fit the particular application at hand, and it is fixed in the optimization process.

We begin by defining the filter error as

$$e(t) = s(t + \alpha) - x(t) \quad (4.3.1)$$

The squared error is then

$$e^2(t) = s^2(t + \alpha) - 2s(t + \alpha)x(t) + x^2(t) \quad (4.3.2)$$

We next write $x(t)$ as a convolution integral

$$\int_{-\infty}^{\infty} g(u)[s(t - u) + n(t - u)] du \quad (4.3.3)$$

*Two-sided Laplace transform theory is used extensively in this section. See the Appendix for a brief review of two-sided transform theory.

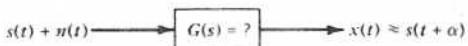


Figure 4.3. Wiener filter problem.

This can then be substituted into Eq. (4.3.2) and both sides averaged to yield*

$$\begin{aligned} E(e^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u)g(v)R_{s+n}(u-v) du dv \\ &\quad - 2 \int_{-\infty}^{\infty} g(u)R_{s+n,s}(\alpha+u)du + R_s(0) \end{aligned} \quad (4.3.4)$$

where

R_s = autocorrelation function of $s(t)$

R_{s+n} = autocorrelation function of $s(t) + n(t)$

$R_{s+n,s}$ = cross-correlation between $s(t) + n(t)$ and $s(t)$

Note that if signal and noise are uncorrelated and have zero means,

$$\left. \begin{array}{l} R_{s+n} = R_s + R_n \\ R_{s+n,s} = R_s \end{array} \right\} \quad (\text{for } s \text{ and } n \text{ uncorrelated}) \quad (4.3.5)$$

We wish to find the function $g(u)$ in Eq. (4.3.4) that minimizes $E(e^2)$. This will be recognized as a problem in calculus of variations (7). Following the usual procedure, we replace $g(u)$ with a perturbed weighting function $g(u) + \varepsilon\eta(u)$ where

$g(u)$ = optimum weighting function. (*Note:* From this point on in the solution, $g(u)$ will denote the *optimal* weighting function.)

$\eta(u)$ = an arbitrary perturbing function.

ε = small perturbation factor such that the perturbed function approaches the optimum one as ε goes to zero.

The optimum and perturbed weighting functions are sketched in Fig. 4.4. Replacing $g(u)$ with $g(u) + \varepsilon\eta(u)$ in Eq. (4.3.4) then leads to

$$\begin{aligned} E(e^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [g(u) + \varepsilon\eta(u)][g(v) + \varepsilon\eta(v)]R_{s+n}(u-v) du dv \\ &\quad - 2 \int_{-\infty}^{\infty} [g(u) + \varepsilon\eta(u)]R_{s+n,s}(\alpha+u)du + R_s(0) \end{aligned} \quad (4.3.6)$$

*We have chosen here to write the mean-square error in terms of the filter weighting function and input autocorrelation functions. In the stationary problem, one can also write $E(e^2)$ in terms of the filter transfer function and input spectral functions, and then proceed with the optimization on that basis (5,6). We have chosen the time-domain approach because it is easily generalized to the nonstationary problem, which is considered in Section 4.4. The frequency-domain approach is not readily generalized.

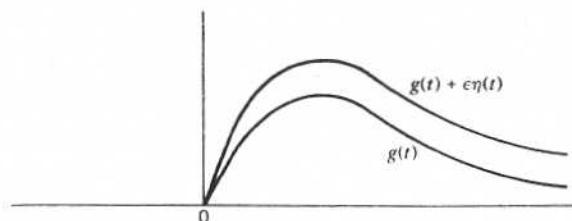


Figure 4.4. Optimal and perturbed weighting functions.

Note that $E(e^2)$ is now a function of ε , and it is to be a minimum when $\varepsilon = 0$. Now, using differential calculus methods, we differentiate $E(e^2)$ with respect to ε and set the result equal to zero for $\varepsilon = 0$. After interchanging dummy variables of integration freely, the result is

$$\int_{-\infty}^{\infty} \eta(\tau)[-R_{s+n,s}(\alpha+\tau) + \int_{-\infty}^{\infty} g(u)R_{s+n}(u-\tau) du] d\tau = 0 \quad (4.3.7)$$

A subtlety in the solution arises at this point; therefore it is convenient to look at the causal and noncausal cases separately.

Noncausal Solution

If we put no constraint on the filter weighting function, we will very likely obtain a $g(u)$ which is nontrivial for negative as well as positive u . This weighting function is noncausal because it requires the filter to "look ahead" of real time and use data that is not yet available. This is, of course, not possible if the filter is operating on line. However, in off-line applications, such as postflight analysis of recorded data, the noncausal solution is possible and very much of interest. Thus, it should not be ignored.

If there are no restrictions on $g(u)$, then, similarly, there are no constraints on the perturbation function $\eta(\tau)$. It is arbitrary for all values of its argument. Thus, if the integral with respect to τ in Eq. (4.3.7) is to be zero, the bracketed term must be zero for all τ . This leads to

$$\int_{-\infty}^{\infty} g(u)R_{s+n}(u-\tau) du = R_{s+n,s}(\alpha+\tau), \quad -\infty < \tau < \infty \quad (4.3.8)$$

This is an integral equation of the first kind, and in this case it can be solved readily using Fourier transform methods. Since R_{s+n} is symmetric, the term on the left side of Eq. (4.3.8) has the exact form of a convolution integral. Therefore, transforming both sides yields

$$G(s)S_{s+n}(s) = S_{s+n,s}(s)e^{as} \quad (4.3.9)$$

or

$$G(s) = \frac{S_{s+n,s}(s)e^{as}}{S_{s+n}(s)} \quad (4.3.10)$$

Remember that the transforms indicated in Eq. (4.3.10) are *two-sided* transforms rather than the usual single-sided transforms. Of course, if we wish to find the weighting function $g(u)$, we simply take the inverse transform of the expression given by Eq. (4.3.10).

The filter mean-square error is given by Eq. (4.3.4). If $g(u)$ is the optimal weighting function satisfying Eq. (4.3.8), the mean-square error equation may be simplified as follows. First write the second term of Eq. (4.3.4) as the sum of two equal terms and combine one of these with the double integral term. After rearranging terms, this leads to

$$\begin{aligned} E(e^2) &= R_s(0) - \int_{-\infty}^{\infty} g(u)R_{s+n,s}(\alpha + u) du \\ &\quad + \int_{-\infty}^{\infty} g(u)[-R_{s+n,s}(\alpha + u) + \int_{-\infty}^{\infty} g(v)R_{s+n}(v - u)dv]du \end{aligned} \quad (4.3.11)$$

The bracketed quantity in Eq. (4.3.11) is zero for optimal $g(v)$ for all u . Therefore the mean-square error is

$$E(e^2) = R_s(0) - \int_{-\infty}^{\infty} g(u)R_{s+n,s}(\alpha + u) du \quad (4.3.12)$$

Example 4.2 Consider the same Markov signal and white noise combination used in Example 4.1. We wish to find the optimal noncausal filter (i.e., $\alpha = 0$). In order to simplify the arithmetic, let $\sigma^2 = \beta = A = 1$. Since the signal and noise are uncorrelated,

$$S_{s+n} = S_s + S_n = \frac{2}{-s^2 + 1} + 1 = \frac{-s^2 + 3}{-s^2 + 1} \quad (4.3.13)$$

$$S_{s+n,s} = S_s = \frac{2}{-s^2 + 1} \quad (4.3.14)$$

$$e^{\alpha s} = 1 \quad (4.3.15)$$

From Eq. (4.3.10) we have

$$G(s) = \frac{2}{\frac{-s^2 + 1}{-s^2 + 3}} = \frac{2}{-s^2 + 3} \quad (4.3.16)$$

Expanding this with a partial fraction expansion yields

$$G(s) = \frac{1/\sqrt{3}}{s + \sqrt{3}} + \frac{1/\sqrt{3}}{-s + \sqrt{3}} \quad (4.3.16)$$

The positive- and negative-time parts of $g(u)$ are given by the first and second terms of Eq. (4.3.16). Thus $g(u)$ is

$$g(u) = \begin{cases} \frac{1}{\sqrt{3}} e^{-\sqrt{3}u}, & u \geq 0 \\ \frac{1}{\sqrt{3}} e^{\sqrt{3}u}, & u < 0 \end{cases} \quad (4.3.17)$$

This is the optimal noncausal weighting function, and it is sketched in Fig. 4.5 along with the intuitive weighting function of Example 4.1 (evaluated for $\sigma^2 = \beta = A = 1$ and $T = 1 + \sqrt{2}$). Note that since the noncausal filter weights both past and future input data, it can afford to have a smaller time constant than the intuitive filter, which is only allowed to weight past input data.

It is also of interest to compare the mean-square errors for the noncausal optimal filter and the parameter-optimized filter of Example 4.1. These may be computed from Eqs. (4.2.6) and (4.3.12) with the result

$$E(e^2)(\text{parameter-optimized}) \approx 0.914$$

$$E(e^2)(\text{noncausal optimal}) \approx 0.577$$

Notice that the noncausal filter has significantly less error than the causal one. Certainly in off-line applications it would be worthwhile implementing the noncausal filter in preference to the intuitive causal one. ■

Causal Solution

The calculus of variations procedure led to Eq. (4.3.7), which is repeated here for convenient reference.

$$\int_{-\infty}^{\infty} \eta(\tau)[-R_{s+n,s}(\alpha + \tau) + \int_{-\infty}^{\infty} g(u)R_{s+n}(u - \tau)du]d\tau = 0 \quad (4.3.7)$$

Recall that $\eta(\tau)$ is an arbitrary perturbing function. If we wish to constrain the filter weighting function to be causal, we must place a similar constraint on $\eta(\tau)$ in the variation. Otherwise, we get the unconstrained (noncausal) solution. Thus, for the causal case we require $\eta(\tau)$ to be zero for negative τ , and allow it to be arbitrary for positive τ . The bracketed quantity in Eq. (4.3.7) then needs to be zero only for positive τ . The zero criterion is satisfied for negative τ by

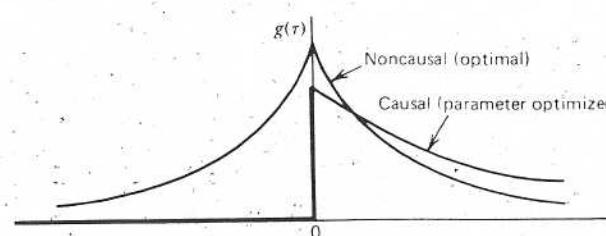


Figure 4.5. Filter weighting functions for optimal noncausal and parameter-optimized causal filters.

virtue of our constraint on $\eta(\tau)$, that is, $\eta(\tau) = 0$ for $\tau < 0$. Therefore, the resulting integral equation is

$$\int_{-\infty}^{\infty} g(u)R_{s+n}(u - \tau)du - R_{s+n,s}(\alpha + \tau) = 0, \quad \tau \geq 0 \quad (4.3.18)$$

Equation (4.3.18) is known as the *Wiener-Hopf equation*, and the fact that it is only valid for $\tau \geq 0$ complicates the solution considerably.

One solution of Eq. (4.3.18) that is based on spectral factorization proceeds as follows. First, replace the right side with an unknown negative-time function $a(\tau)$ [i.e., $a(\tau)$ is unknown for negative time, but is known to be zero for positive time]. Eq. (4.3.18) can then be written as

$$\int_{-\infty}^{\infty} g(u)R_{s+n}(u - \tau)du - R_{s+n,s}(\alpha + \tau) = a(\tau), \quad -\infty < \tau < \infty \quad (4.3.19)$$

Transforming both sides of Eq. (4.3.19) then yields

$$G(s)S_{s+n}(s) - S_{s+n,s}(s)e^{\alpha s} = A(s) \quad (4.3.20)$$

Next, use spectral factorization on S_{s+n} and group terms as follows:

$$[G(s)S_{s+n}^+(s)]S_{s+n}^-(s) - S_{s+n,s}(s)e^{\alpha s} = A(s)$$

or

$$G(s)S_{s+n}^-(s) = \frac{A(s)}{S_{s+n}^-(s)} + \frac{S_{s+n,s}(s)e^{\alpha s}}{S_{s+n}^-(s)} \quad (4.3.21)$$

In Eq. (4.3.21), the “super +” indicates the factored part of the spectral function that has all its poles and zeros in the left half-plane. Similarly, the poles and zeros of S_{s+n}^- are mirror images of those of S_{s+n}^+ . We note here that $g(u)$ is a stable positive-time function; therefore $G(s)$ will have its poles in the left half plane. Thus, $G(s)S_{s+n}^-(s)$ will have all its poles in the left half plane, and it will be the transform of a positive-time function. Similarly, $A(s)$ is the transform of a negative-time function, so its poles will be in the right half-plane. Also, both the zeros and poles of $S_{s+n}^-(s)$ are in the right half-plane. Hence, the three terms of Eq. (4.3.21) translate into words as

$$\begin{bmatrix} \text{Positive-time} \\ \text{function} \end{bmatrix} = \begin{bmatrix} \text{Negative-time} \\ \text{function} \end{bmatrix} + \begin{bmatrix} \text{Both positive-} \\ \text{and negative-time} \\ \text{function} \end{bmatrix}$$

Equating positive-time parts on both sides of Eq. (4.3.21) then leads to

$$G(s)S_{s+n}^-(s) = \text{positive-time part of } \frac{S_{s+n,s}(s)e^{\alpha s}}{S_{s+n}^-(s)}$$

or

$$G(s) = \frac{1}{S_{s+n}^-(s)} \left[\text{positive-time part of } \frac{S_{s+n,s}(s)e^{\alpha s}}{S_{s+n}^-(s)} \right] \quad (4.3.22)$$

The bracketed term of Eq. (4.3.22) can be interpreted as follows. First, find the inverse transform of $S_{s+n,s}(s)/S_{s+n}^-(s)$. This will normally be nontrivial for both positive and negative time. Next, translate the time function an amount α . (This accounts for $e^{\alpha s}$). Finally, take the ordinary *single-sided* Laplace transform of the shifted time function, and this will be the bracketed quantity in Eq. (4.3.22). Two examples should be helpful at this point.

Example 4.3 Consider the same Markov signal and white noise combination used in Examples 4.1 and 4.2. Again we let $\sigma^2 = \beta = A = 1$ to simplify the arithmetic and, in this example, we are looking for the optimal causal solution. Since the signal and noise are assumed to be uncorrelated, the needed spectral functions are

$$S_{s+n} = S_s + S_n = \frac{2}{-s^2 + 1} + 1 = \frac{-s^2 + 3}{-s^2 + 1} \quad (4.3.23)$$

$$S_{s+n,s} = S_s = \frac{2}{-s^2 + 1} \quad (4.3.24)$$

Also, since the prediction time α is assumed to be zero,

$$e^{\alpha s} = 1 \quad (4.3.25)$$

First, we factor S_{s+n} :

$$S_{s+n} = S_{s+n}^+ S_{s+n}^- = \left[\frac{s + \sqrt{3}}{s + 1} \right] \left[\frac{-s + \sqrt{3}}{-s + 1} \right] \quad (4.3.26)$$

Next, we form the $S_{s+n,s}/S_{s+n}^-$ function

$$\frac{S_{s+n,s}}{S_{s+n}^-} = \frac{\frac{2}{-s^2 + 1}}{\frac{-s + \sqrt{3}}{-s + 1}} = \frac{2}{(-s + \sqrt{3})(s + 1)} \quad (4.3.27)$$

This, in turn, can be expanded in terms of a partial fraction expansion

$$\frac{S_{s+n,s}}{S_{s+n}^-} = \frac{\sqrt{3} - 1}{s + 1} + \frac{\sqrt{3} - 1}{-s + \sqrt{3}} \quad (4.3.28)$$

Clearly, the first term of Eq. (4.3.28) is the positive-time part. Therefore, $G(s)$ as given by Eq. (4.3.22) is

$$G(s) = \frac{1}{s + \sqrt{3}} \cdot \frac{\sqrt{3} - 1}{s + 1} = \frac{\sqrt{3} - 1}{s + \sqrt{3}} \quad (4.3.29)$$

Or, in terms of the filter weighting function,

$$g(t) = \begin{cases} (\sqrt{3} - 1)e^{-\sqrt{3}t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (4.3.30)$$

Table 4.1 Comparison of Results of Examples 4.1, 4.2, and 4.3

Type Of Solution	Filter Transfer Function	Weighting Function	Mean-Square Error
Single parameter optimization	$G(s) = \frac{1}{1 + (1 + \sqrt{2})s}$	$g(t) = \begin{cases} \frac{1}{1 + \sqrt{2}} e^{-(\sqrt{2}-1)t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$.914
Causal Wiener filter	$G(s) = \frac{\sqrt{3} - 1}{s + \sqrt{3}}$	$g(t) = \begin{cases} (\sqrt{3} - 1)e^{-\sqrt{3}t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$.732
Noncausal Wiener filter	$G(s) = \frac{2}{(s + \sqrt{3})(-s + \sqrt{3})}$	$g(t) = \begin{cases} \frac{1}{\sqrt{3}} e^{-\sqrt{3}t}, & t \geq 0 \\ \frac{1}{\sqrt{3}} e^{\sqrt{3}t}, & t < 0 \end{cases}$.577

As before, the mean-square error of the filter can be computed using Eq. (4.3.12). The result is

$$E(e^2) = .732 \quad (4.3.31)$$

We have now examined three different optimization approaches for the same signal-plus-noise situation. A comparison of the results is shown in Table 4.1, with the most restrictive filter being listed first and the least restrictive one listed last. As should be expected, the mean-square error decreases with each successive relaxation of the constraints on the choice of transfer function. The linear constraint is, of course, present in all three solutions. We will see in later chapters that this is not as serious as one might think at first glance. The explanation of this, though, will be deferred until Chapter 5.

Example 4.4 There is a classical problem in random process theory known as the *pure prediction* problem. Here we assume the additive noise corrupting the signal is zero, and we pose the problem of looking ahead and finding the best estimate of the signal α units ahead of the present time t . To demonstrate the applicability of Wiener filter theory to this problem, let the signal be Markov with a known autocorrelation function

$$R_s(\tau) = \sigma^2 e^{-\beta|\tau|} \quad \text{or} \quad S_s(s) = \frac{2\sigma^2\beta}{-s^2 + \beta^2} \quad (4.3.32)$$

Also,

$$R_n(\tau) = 0$$

4.3 STATIONARY OPTIMIZATION PROBLEM—WEIGHTING FUNCTION APPROACH 159

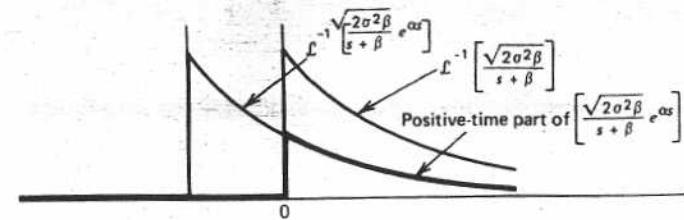


Figure 4.6. The shifted time functions for Example 4.4.

We first factor $S_{s+n}(s)$.

$$S_{s+n} = S_s = S_s^+ \cdot S_s^- = \frac{\sqrt{2\sigma^2\beta}}{s + \beta} \cdot \frac{\sqrt{2\sigma^2\beta}}{-s + \beta} \quad (4.3.33)$$

Next, we form the $S_{s+n,s}/S_{s+n}^-$ function

$$\frac{S_{s+n,s}}{S_{s+n}^-} = \frac{S_s}{S_s^-} = S_s^+ = \frac{\sqrt{2\sigma^2\beta}}{s + \beta} \quad (4.3.34)$$

In this problem α is not zero, and hence we must first multiply Eq. (4.3.34) by $e^{\alpha s}$ and then find the positive-time part of the result. This is readily accomplished by appropriate shifting in the time domain as shown in Fig. 4.6. Finally, substituting the proper positive-time part into Eq. (4.3.22) yields

$$G(s) = \frac{1}{\sqrt{2\sigma^2\beta}} \cdot e^{-\alpha\beta} \frac{\sqrt{2\sigma^2\beta}}{s + \beta} = e^{-\alpha\beta} \quad (4.3.35)$$

Or, the corresponding weighting function is

$$g(t) = e^{-\alpha\beta} \delta(t) \quad (4.3.36)$$

The Wiener solution says that the best one can hope to do (in the least-squares sense) is to multiply the present value of the input by an attenuation factor $e^{-\alpha\beta}$, and this will yield the best estimate of the process α units ahead of the present time. Observe that the predictive estimate is dependent only on the present value of the input signal and not on its past history. If the Markov signal under consideration is also Gaussian, then the predictive estimate given by Wiener theory is also identical to the conditional mean of the process at the predicted time, conditioned on all the prior input data. The use of the term Markov for a process with an exponential autocorrelation function should now be apparent. This will be expanded on later in Chapter 5. ■

4.4 THE NONSTATIONARY PROBLEM

In the preceding discussion it was assumed that the filter was "turned on" at $t = -\infty$, which made the entire past history of $s(t) + n(t)$ available for weighting. This led to the steady-state or stationary solution. In the transient or nonstationary problem we consider the signal and noise to be covariance stationary processes with known spectral characteristics as before, but we consider the input to be applied at $t = 0$ rather than $-\infty$. This is indicated in Fig. 4.7 as a switching operation at the input. We assume zero initial conditions; hence the filter output can be written as

$$x(t) = \int_0^t g(\tau)[s(t - \tau) + n(t - \tau)] d\tau, \quad t \geq 0 \quad (4.4.1)$$

As before, the ideal output would be $s(t + \alpha)$. Thus, the filter error is

$$e(t) = s(t + \alpha) - x(t) \quad (4.4.2)$$

Substituting Eq. (4.4.1) into Eq. (4.4.2), squaring, and then taking the expectation of both sides yields

$$\begin{aligned} E(e^2) &= \int_0^t \int_0^t g(u)g(v)R_{s+n}(u - v) du dv \\ &\quad - 2 \int_0^t g(u)R_{s+n,s}(\alpha + u) du + R_s(0) \end{aligned} \quad (4.4.3)$$

The problem is to choose $g(u)$ such as to minimize $E(e^2)$. The variational procedure to be followed is essentially the same as for the stationary case, so it will not be repeated. The only difference is in the limits on the integration, which, in turn, traces back to the expression for $x(t)$ given by Eq. (4.4.1). This is an important difference, though. We need not worry about placing a causality constraint on $g(\tau)$ or its perturbation, because the range of integration is limited to be from 0 to t . We can arbitrarily truncate $g(\tau)$ to zero outside this range.

As before, the variational procedure leads to an integral equation in $g(\tau)$. For the nonstationary problem, it is

$$\int_0^t g(u)R_{s+n}(\tau - u) du = R_{s+n,s}(\alpha + \tau), \quad 0 \leq \tau \leq t \quad (4.4.4)$$

This equation is similar to the Wiener-Hopf equation except for the limits on the integral and the range of τ for which the equation is valid. These seemingly

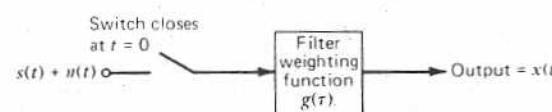


Figure 4.7. Block diagram for the nonstationary problem.

small differences complicate the solution even further, though. The spectral factorization method cannot be applied to Eq. (4.4.4) because the range of τ is finite rather than semi-infinite as in the Wiener-Hopf equation. Equation (4.4.4) is amenable to solution, though, by another method that involves transforming the integral equation to a differential equation and then solving the differential equation for $g(\tau)$ (8). This technique will only work for the case where the spectral function for $s(t) + n(t)$ is rational, that is, a ratio of polynomials in s^2 . However, we note that this same restriction is necessary for spectral factorization, and therefore only limited types of spectral situations can be handled for either the stationary or nonstationary case.

The solution of Eq. (4.4.4) proceeds as follows. We first write R_{s+n} as a Fourier integral

$$\int_0^t g(u) \left[\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} S_{s+n}(s) e^{s(\tau-u)} ds \right] du = R_{s+n,s}(\alpha + \tau) \quad (4.4.5)$$

Next, we write $S_{s+n}(s)$ as a ratio of polynomials in s^2

$$S_{s+n}(s) = \frac{N(s^2)}{D(s^2)} \quad (4.4.6)$$

Now we note that operating on $e^{s(\tau-u)}$ with the differential operator $D(d^2/d\tau^2)$ will generate $D(s^2)e^{s(\tau-u)}$.* Thus the denominator of S_{s+n} in Eq. (4.4.5) can be cancelled by operating on both sides of the equation with $D(d^2/d\tau^2)$. Similarly, algebraic multiplication within the integral by $N(s^2)$ is equivalent to operating in front of the first integral with $N(d^2/d\tau^2)$. Inserting these equivalent operations in Eq. (4.4.5) yields

$$N \left(\frac{d^2}{d\tau^2} \right) \int_0^t g(u) \left[\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} e^{s(\tau-u)} ds \right] du = D \left(\frac{d^2}{d\tau^2} \right) R_{s+n,s}(\alpha + \tau) \quad (4.4.7)$$

We now note that the Fourier integral in brackets in Eq. (4.4.7) is just the Dirac delta function $\delta(\tau - u)$. Inserting this and using the sifting property of the impulse function leads to

$$N \left(\frac{d^2}{d\tau^2} \right) g(\tau) = D \left(\frac{d^2}{d\tau^2} \right) R_{s+n,s}(\alpha + \tau), \quad 0 < \tau < t \quad (4.4.8)$$

We now have a differential equation in $g(\tau)$ rather than an integral equation. Furthermore, it is a linear differential equation with constant coefficients, and the solution of this type of equation is well known. Before proceeding, note that the interval on τ in Eq. (4.4.8) is the open interval $(0, t)$ rather than the closed interval $[0, t]$ associated with the integral equation, Eq. (4.4.4). This is

* Remember that D refers to the polynomial in the denominator of $S_{s+n}(s)$ and not the "D-operator" used in differential equation theory. That is, $D(d^2/d\tau^2)$ is $D(s^2)$ with s^2 replaced by $d^2/d\tau^2$. Similarly, $N(d^2/d\tau^2)$ is the numerator polynomial $N(s^2)$ with s^2 replaced with $d^2/d\tau^2$.

intentional and arises because of the problem of continuity and differentiation at the end points. In other words, we may only safely assume that the differential equation is valid in the interior region of the interval.

The solution of Eq. (4.4.8) will, of course, contain arbitrary constants of integration. Furthermore, because of the end-point problem, impulse functions with undetermined amplitudes must be added at $\tau = 0+$ and $\tau = t-$. (The + and - indicate that the impulses are placed at the inside edges of the interval.) The rule for adding the impulses is as follows: If the order of $D(s^2)$ is:

1. The same as $N(s^2)$, add no impulses.
2. If it is two greater than $N(s^2)$, add simple impulses.
3. If it is four greater than $N(s^2)$, add simple impulses plus doublet impulses.
4. Etc.

The unknown coefficients in the general solution are evaluated by substituting the assumed solution into the original integral equation and demanding equality on both sides of the equation. This is much the same as using the initial conditions to evaluate the constants of integration in the usual initial-condition problem.

Remember that the procedure just described is highly specialized and only applies to the case where $S_{s+n}(s)$ can be written as a ratio of polynomials in s^2 . The justification of the procedure in any particular case lies in the final substitution of the solution into the integral equation. We can ask no more of the solution than to satisfy the original integral equation.

Example 4.5 We again look at the situation where the signal is Markov and the noise is white. The additive combination forms the input which is applied to the filter at $t = 0$. Let

$$R_s(\tau) = e^{-|\tau|} \quad \text{or} \quad S_s(s) = \frac{2}{-s^2 + 1} \quad (4.4.9)$$

$$R_n(\tau) = \delta(\tau) \quad \text{or} \quad S_n(s) = 1 \quad (4.4.10)$$

If we assume the signal and noise are independent,

$$S_{s+n}(s) = \frac{2}{-s^2 + 1} + 1 = \frac{-s^2 + 3}{-s^2 + 1} = \frac{N(s^2)}{D(s^2)} \quad (4.4.11)$$

Also, if $\alpha = 0$

$$R_{s+n,s}(\alpha + \tau) = R_s(\tau) = e^{-\tau}, \quad \tau \geq 0 \quad (4.4.12)$$

Note that the absolute magnitude signs around τ may be dropped because τ is always positive. Using the polynomials $N(s^2)$ and $D(s^2)$ as given by Eq. (4.4.11), the differential equation (4.4.8) becomes

$$\left(-\frac{d^2}{d\tau^2} + 3 \right) g(\tau) = \left(-\frac{d^2}{d\tau^2} + 1 \right) e^{-\tau} \quad (4.4.13)$$

and this reduces to

$$-\frac{d^2g(\tau)}{d\tau^2} + 3g(\tau) = 0 \quad (4.4.14)$$

The general solution of this equation is recognized to be

$$g(\tau) = ae^{-\sqrt{3}\tau} + be^{\sqrt{3}\tau} \quad (4.4.15)$$

We do not need to add impulses in this case because $N(s^2)$ and $D(s^2)$ are both second order. Thus, we know the filter weighting function is of the form given by Eq. (4.4.15) without additional impulses. The a and b coefficients may be evaluated by substituting the known form of solution [i.e., Eq. (4.4.15)] into the original integral equation, Eq. (4.4.4), and then choosing a and b such that the resulting functions on left and right sides of the equation are identical functions of τ . This is straightforward, but considerable algebra is involved, which will be omitted. The end result is

$$a(t) = \frac{2(\sqrt{3} + 1)e^{\sqrt{3}t}}{(\sqrt{3} + 1)^2 e^{\sqrt{3}t} - (-\sqrt{3} + 1)^2 e^{-\sqrt{3}t}} \quad (4.4.16)$$

$$b(t) = \frac{-2(-\sqrt{3} + 1)e^{-\sqrt{3}t}}{(\sqrt{3} + 1)^2 e^{\sqrt{3}t} - (-\sqrt{3} + 1)^2 e^{-\sqrt{3}t}} \quad (4.4.17)$$

Note that the "constants" are functions of the running time variable t . The final solution for the weighting function is then

$$g(\tau; t) = a(t)e^{-\sqrt{3}\tau} + b(t)e^{\sqrt{3}\tau} \quad (4.4.18)$$

where $a(t)$ and $b(t)$ are given by Eqs. (3.4.16) and (3.4.17). The semicolon in $g(\tau; t)$ is used to emphasize the fact that τ is the usual age variable in the weighting function and t is just a parameter. The resulting filter is, of course, a time-variable filter.

It is readily verified that as t approaches ∞ , the solution for $g(\tau)$ becomes

$$g(\tau) = (\sqrt{3} - 1)e^{-\sqrt{3}\tau} \quad (4.4.19)$$

As should be expected, this is the same steady-state solution that was obtained previously using spectral factorization methods. It is of interest to note that the differential-equation approach provides an alternative method of solving the stationary problem. ■

In Example 4.5 it is worth noting that the running time variable t came into the weighting-function solution naturally (i.e., without any conscious effort) because we chose to write the superposition integral in the form

$$x(t) = \int_0^t g(\tau; t)f(t - \tau) d\tau \quad (4.4.20)$$

The other form, which is equally valid, and sometimes preferred in books on linear systems theory (9), is

$$x(t) = \int_0^t h(t,\tau)f(\tau) d\tau \quad (4.4.21)$$

In Eq. (4.4.21), $h(t,\tau)$ has the physical meaning of the system response to a unit impulse applied at time τ . The relationship between the impulsive response and weighting function is obtained by making the appropriate change of variable in either Eq. (4.4.20) or Eq. (4.4.21) and then comparing the two integrals. The result is

$$h(t,\tau) = g(t - \tau; t) \quad (4.4.22)$$

4.5 ORTHOGONALITY

It was shown in Section 4.4 that the filter weighting function that minimizes the mean-square error must satisfy the integral equation

$$\int_0^t g(u)R_{s+n}(\tau - u) du = R_{s+n,s}(\alpha + \tau), \quad 0 \leq \tau \leq t \quad (4.5.1)$$

Also, the filter error is given by

$$e(t) = s(t + \alpha) - x(t) = s(t + \alpha) - \int_0^t g(u)[s(t - u) + n(t - u)] du \quad (4.5.2)$$

We wish to examine the expectation of the product of the filter error at time t and input at some time t_1 where $0 \leq t_1 \leq t$. Let the input be denoted as $z(t)$. Then

$$z(t_1) = s(t_1) + n(t_1) \quad (4.5.3)$$

and

$$E[z(t_1)e(t)] = E \left[\left\{ s(t_1) + n(t_1) \right\} \left\{ s(t + \alpha) - \int_0^t g(u)[s(t - u) + n(t - u)] du \right\} \right] \quad (4.5.4)$$

Moving $s(t_1) + n(t_1)$ inside the integration and carrying out the expectation operation yields

$$E[z(t_1)e(t)] = R_{s+n,s}(t - t_1 + \alpha) - \int_0^t g(u)R_{s+n}(t - t_1 - u) du \quad (4.5.5)$$

However, $g(u)$ must satisfy the integral equation, (4.5.1). Thus, the above must be zero for $0 \leq (t - t_1) \leq t$. Since the time t_1 was assumed to lie between 0 and t , this is equivalent to saying

$$E[z(t_1)e(t)] = 0, \quad 0 \leq t_1 \leq t \quad (4.5.6)$$

If the expectation of the product of two random variables is zero, the variables are said to be *orthogonal*. Equation (4.5.6) states that the filter error at the current time t is not only orthogonal to the input at the same time t , but it is also orthogonal to the input evaluated at any previous time during the past history of the filter operation. This is a consequence of minimizing the mean-square error. Furthermore, it should be apparent from the derivation that the argument can be reversed. That is, if we begin by assuming that $e(t)$ is orthogonal to $z(t_1)$, we can then conclude that the integral equation, (4.5.1), is satisfied. It is important to recognize this equivalence because some authors, especially in recent literature, prefer to begin their optimality arguments with the orthogonality relationship rather than the minimization of the mean-square error (2,3,8).

4.6 COMPLEMENTARY FILTER

Applications of Wiener filter theory are not as commonplace as one might expect. Perhaps one reason for this is that Wiener theory demands that the signal be noiselike in character. In the usual communication problem, this is not the case. The signal usually has at least some deterministic structure, and it is often not reasonable to assume it to be completely random. Thus, the typical filtering problem encountered in communication engineering simply does not fit the Wiener mold. There is an instrumentation application, though, where Wiener methods have been used extensively. In this application, redundant measurements of the same signal are available, and the problem is to combine all the information in such a way as to minimize the instrumentation errors. In order to keep the discussion as simple as possible, the two-input case will be considered first.

Consider the general problem of combining two independent noisy measurements of the same signal as depicted in Fig. 4.8. In the context of instrumentation, the measurements might come from two completely different types of instruments, each with its own particular error characteristic. We wish to blend the two measurements together in such a way as to eliminate as much of the error as possible. If the signal $s(t)$ is noiselike, Wiener methods may be used to determine the transfer functions $G_1(s)$ and $G_2(s)$ that minimize the mean-square error. This is the two-input Wiener problem, and more will be said of it in the next section. However, more often than not the signal may not be

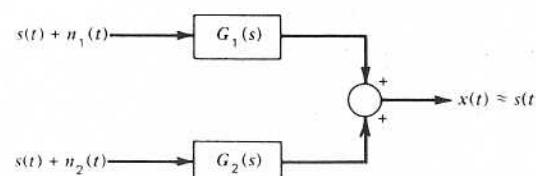


Figure 4.8 General two-input Wiener problem.

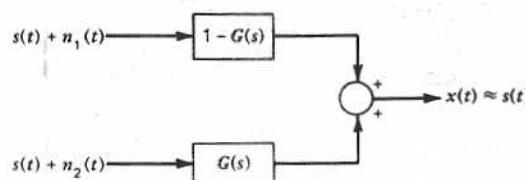


Figure 4.9. Complementary filter.

properly modeled as a random process with known spectral characteristics. For example, if $s(t)$ represents the position of an airplane in flight along a prescribed air route, certainly the signal is not random. Furthermore, in this case we would not want to delay or distort the signal in any way in the process of filtering the measurement errors.

A method of filtering the noise without distorting the signal is shown in Fig. 4.9. From the block diagram, it should be apparent that the Laplace transform of the output may be written as

$$X(s) = \underbrace{S(s)}_{\text{Signal term}} + \underbrace{N_1(s)[1 - G(s)] + N_2(s)G(s)}_{\text{Noise term}} \quad (4.6.1)$$

Clearly, the signal term $S(s)$ is not affected by our choice of $G(s)$ in any way. On the other hand, the two noise inputs are modified by the complementary transfer functions $[1 - G(s)]$ and $G(s)$. If the two noises have complementary spectral characteristics, $G(s)$ may be chosen to mitigate the noise in both channels. For example, if n_1 is predominantly low-frequency noise and n_2 high-frequency, then choosing $G(s)$ to be a low-pass filter will automatically attenuate n_1 as well as n_2 .

We now note that the noise term in Eq. (4.6.1) has the same form as seen before [Eq. (4.2.4)] except for the sign on the $N_2(s)$ term. Thus, we would expect to be able to use Wiener methods in minimizing this term. This is perhaps even more evident from Fig. 4.10. It can be easily verified that the input-output relationships are identical for the systems of Figs. 4.9 and 4.10, and thus they are equivalent. From Fig. 4.10 we see that the purpose of the filter $G(s)$ is to give the best possible estimate of $n_1(t)$, and this, in turn, is subtracted from $s(t) + n_1(t)$ in order to give an improved estimate of $s(t)$. The input to $G(s)$

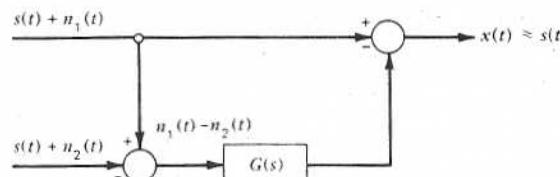


Figure 4.10. Differencing and feedforward configuration for a complementary filter.

is $n_1(t) - n_2(t)$, and hence the filter must separate one *noiselike* signal from another. Clearly, if we let $n_1(t)$ play the role of signal and $-n_2(t)$ the role of the noise, this problem fits the Wiener theory perfectly. An example will illustrate an engineering application of the complementary-filter method of combining redundant measurement data.

Example 4.6* In the standard aircraft instrument landing system (ILS), the aircraft descends along a radio beam that provides the pilot with steering signals proportional to the plane's deviation from the desired glide path. The pilot uses these to correct the plane's course in order to follow the beam as closely as possible. In automating the system, say in the lateral channel, it is desirable to have a rate signal to aid in stabilizing the closed-loop control system. One such signal can be obtained by differentiating (approximately) the lateral-deviation radio signal. However, this gives rise to high-frequency noise in the rate signal. Another completely independent lateral rate signal may be obtained from the aircraft heading. If the aircraft speed is constant, an incremental change in heading, when multiplied by speed, will yield a corresponding lateral rate change. This signal, however, is also contaminated with error, the major source in this case being wind changes that occur during the plane's descent. These errors are primarily low-frequency in nature. Thus, we have the ideal setting here for a complementary filter. We have two independent measurements of the same quantity, one contaminated with high-frequency noise and the other with low-frequency noise.

At this point, we assume the following spectral characteristics for the two noises:

1. *Heading-derived rate signal.* Markov noise (primarily low-frequency):

$$S_{n_1}(s) = \frac{2\sigma_1^2\omega_1}{-s^2 + \omega_1^2} \quad (4.6.2)$$

2. *Radio-derived rate signal.* Differentiated Markov noise (primarily high-frequency):

$$S_{n_2}(s) = \frac{2\sigma_2^2\omega_2(-s^2)}{-s^2 + \omega_2^2} \quad (4.6.3)$$

We can now take either of two approaches. The filter form may be assumed and then it can be optimized with respect to a parameter (e.g., the time constant); or the more general Wiener approach may be used, treating n_1 as the signal and n_2 as the noise. We take the latter approach and, in order to keep the algebra

* This example is a simplified version of a system described in a patent granted to W. H. Wirkler (10). Even though Wirkler did not use the term complementary filter in his disclosure, the principle is there. The system patented by Wirkler (with some modifications) was described later by Anderson and Fritze (11), and they introduced the term complementary filter in their paper.

reasonable, we assume plausible values for the various parameters. Let

$$\begin{aligned}\sigma_1 &= 2 \text{ m/sec} \\ \omega_1 &= 1/20 \text{ rad/sec}\end{aligned}\left\{\begin{array}{l}\text{Heading-derived} \\ \text{velocity noise}\end{array}\right.$$

$$\begin{aligned}\sigma_2 &= 20 \text{ m} \\ \omega_2 &= 1/2 \text{ rad/sec}\end{aligned}\left\{\begin{array}{l}\text{Radio-derived} \\ \text{position noise}\end{array}\right.$$

$$\text{Then } S_{n_1+n_2}(s) = \frac{2\sigma_1^2\omega_1}{-s^2 + \omega_1^2} + \frac{2\sigma_2^2\omega_2(-s^2)}{-s^2 + \omega_2^2} \quad (4.6.4)$$

$$\begin{aligned}&= 20 \frac{(s^2 + .187s + .0158)}{(s + .05)(s + .5)} \\ &\quad \cdot 20 \frac{(-s)^2 + .187(-s) + .0158}{(-s + .05)(-s + .5)} \quad (4.6.5)\end{aligned}$$

Spectral factorization of $S_{n_1+n_2}$ is obvious from the way Eq. (4.6.5) has been written. We can now form $S_{n_1}/S_{n_1+n_2}$ as

$$\begin{aligned}\frac{S_{n_1}(s)}{S_{n_1+n_2}(s)} &= \frac{.4}{\frac{(s + .05)(-s + .05)}{(-s)^2 + .187(-s) + .0158}} \\ &= .02 \frac{-s + .5}{(s + .05)[(-s)^2 + .187(-s) + .0158]} \quad (4.6.6)\end{aligned}$$

Clearly, the positive-time part of this will contain only a $K/(s + .05)$ term. Evaluating this yields

$$\text{Positive-time part of } \frac{S_{n_1}}{S_{n_1+n_2}} = \frac{.3976}{s + .05} \quad (4.6.7)$$

The transfer function $G(s)$ is then given by Eq. (4.3.22):

$$\begin{aligned}G(s) &= \frac{1}{20} \frac{.3976}{\frac{s^2 + .187s + .0158}{(s + .05)(s + .5)}} s + .05 \\ &= .0199 \frac{s + .5}{s^2 + .187s + .0158} \quad (4.6.8)\end{aligned}$$

This is obviously a low-pass filter with a break point of about .1 rad/sec. It is interesting to note that the zero-frequency response of $G(s)$ is about .63 rather than unity, as might be expected intuitively. This means that even at very low frequencies the heading-derived signal will be given significant weight—.37 relative to the weight of .63 given the radio signal. Obviously, this will depend on the numerical values assigned to the various parameters. However, in any event, optimal filter theory says the heading-derived signal has at least some

value at low frequencies. On the other hand, the weight shifts entirely to the heading signal at high frequencies. ■

The principle of complementary filtering is easily extended to the case of more than two signals. All one has to do is let one of the transfer characteristics be the complement of the sum of the others. For example, for the three-input case, let

$G_1(s)$ = transfer function for noisy measurement #1

$G_2(s)$ = transfer function for noisy measurement #2

$1 - G_1(s) - G_2(s)$ = transfer function for noisy measurement #3

Then the signal component passes through the system undistorted, and one chooses $G_1(s)$ and $G_2(s)$ such as to give the best suppression of the noise. The problem of determining the optimal $G_1(s)$ and $G_2(s)$ is a multiple-input Wiener problem and this will be discussed later.

It is important to note that the choice of complementary filter transfer function does not depend on any prior assumptions about the *signal* structure. It can be either noiselike or deterministic, and the complementary feature assures that the signal will not be distorted in any way by the filtering action. Thus, philosophically, it is a "safe" design and is particularly applicable in situations where the designer wants a filter that will cope reasonably well with statistically unusual situations without giving disastrously large errors. A strict Wiener design with no complementary constraint is, of course, chosen to minimize the *average squared error*. So, in the unusual situation the error may be quite large. This can be disastrous in some applications.

4.7 MULTIPLE-INPUT WIENER FILTER PROBLEM

Figure 4.11 shows the general two-input Wiener problem which will be used to illustrate the multiple-input problem. Since the extension to more than two inputs is fairly obvious, we confine our attention to the two-input case. We assume that $s(t)$ as well as $n_1(t)$ and $n_2(t)$ are stationary random processes with known spectral characteristics, and in this case, we do *not* wish to constrain

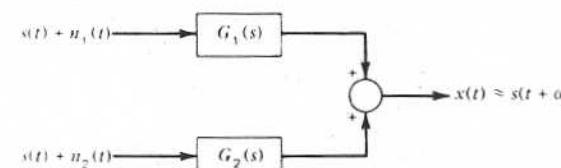


Figure 4.11. General two-input Wiener problem.

one transfer function to be the complement of the other. The problem is to find the causal transfer functions $G_1(s)$ and $G_2(s)$ that will minimize the mean-square error.

Since the solution for the two-input case proceeds much the same as for the single-input problem, some of the details will be omitted. Also, in order to keep the solution as simple as possible, the signal and noises will be assumed to be mutually uncorrelated. We first write the expression for the error as

$$\begin{aligned} e(t) &= s(t + \alpha) - x(t) \\ &= s(t + \alpha) - \int_{-\infty}^{\infty} g_1(\tau) [s(t - \tau) + n_1(t - \tau)] d\tau \\ &\quad - \int_{-\infty}^{\infty} g_2(\tau) [s(t - \tau) + n_2(t - \tau)] d\tau \end{aligned} \quad (4.7.1)$$

The expectation of the square of $e(t)$ is then formed, and this will involve the two weighting functions $g_1(\tau)$ and $g_2(\tau)$. Next, following the usual calculus-of-variations procedure, g_1 and g_2 are replaced with perturbed weighting functions $g_1(\tau) + \epsilon_1 \eta_1(\tau)$ and $g_2(\tau) + \epsilon_2 \eta_2(\tau)$. This allows us to consider independent variations of both weighting functions by varying ϵ_1 , ϵ_2 . After the indicated substitution, the mean-square error becomes a function of ϵ_1 and ϵ_2 . The partial derivatives of $E(e^2)$ with respect to ϵ_1 and ϵ_2 may then be formed and set equal to zero for $\epsilon_1 = \epsilon_2 = 0$. Finally, noting that $\eta_1(\tau)$ and $\eta_2(\tau)$ are arbitrary for $\tau \geq 0$ leads to the following pair of integral equations.

$$\left. \begin{aligned} \int_{-\infty}^{\infty} g_1(u) R_{s+n_1}(\tau - u) du \\ + \int_{-\infty}^{\infty} g_2(u) R_s(\tau - u) du - R_s(\tau + \alpha) = 0 \\ \int_{-\infty}^{\infty} g_1(u) R_s(\tau - u) du \\ + \int_{-\infty}^{\infty} g_2(u) R_{s+n_2}(\tau - u) du - R_s(\tau + \alpha) = 0 \end{aligned} \right\} \tau \geq 0 \quad (4.7.2)$$

Note these are *coupled* integral equations and must be solved simultaneously. Also, because of the causality constraint on g_1 and g_2 , the equations are only valid for $\tau \geq 0$.

The solution of Eq. (4.7.2) is not as straightforward as one might think at first glance. If we first transform the equations and then try using spectral factorization as before, we find that the positive- and negative-time parts are not readily separated and identified. Thus, we must look for another approach. A method of solution that is somewhat cumbersome, but workable, proceeds as follows. First, use differential operators similar to those used in Section 4.4 to convert the two integral equations into differential equations. These equations may then be solved using standard methods of ordinary differential equations.

Impulses may then be added at the end points, if needed, and the general solution may then be substituted into the integral equations to determine the undetermined coefficients. This is a long and laborious method of solution, though. An alternative approach, which is less general but may be used in simple situations, is given by Brown and Nilsson (12). This method is summarized in Problem 4.10. Also, we will see later in Chapter 7 that multiple-input problems may be solved using continuous Kalman filter methods.

As mentioned in Section 4.6, we often wish to use the complementary constraint in instrumentation applications where redundant measurements are available. The complementary constraint is especially important in those situations where we want to avoid making questionable assumptions about the nature of the signal. If there are three or more redundant measurements, the complementary filter problem always reduces to a general multiple-input Wiener problem of one less order than the original problem (see Problem 4.16). For example, if three measurements are available, the problem reduces to the general two-input Wiener problem just considered. In effect, one of the noises, say n_1 , plays the role of signal and the problem is to separate it from the other two noises n_2 and n_3 . Once this is done, the filter's best estimate of n_1 may be subtracted from the first noisy measurement $s + n_1$ to yield an improved estimate of the signal s .

4.8 WEIGHTING FUNCTION APPROACH TO THE SAMPLED-DATA FILTER

The Wiener approach to least-squares filtering is basically a weighting function approach. When viewed this way, the basic problem always reduces to: How should the past history of the input be weighted in order to yield the present best estimate of the variable of interest? It is instructive to see how this approach extends to the discrete-measurement situation.

Consider the filter input to be a sequence of discrete noisy measurements z_1, z_2, \dots, z_n as shown in Fig. 4.12. These are additive combinations of signal and noise; hence $z_1 = s_1 + n_1$, $z_2 = s_2 + n_2$, etc. As before, we denote the filter output as x , and therefore the corresponding samples of the output are x_1, x_2, \dots

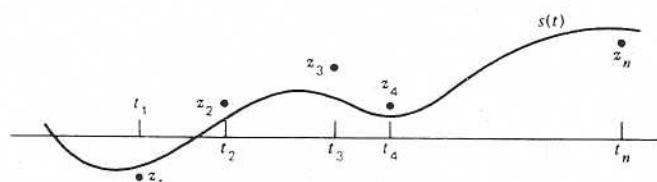


Figure 4.12. Discrete measurement situation.

\dots, x_n . We now write the output at time t_n as a general linear combination of the past measurements

$$x_n = k_1 z_1 + k_2 z_2 + \dots + k_n z_n \quad (4.8.1)$$

The filter error may then be written as

$$\begin{aligned} e_n &= s_n - x_n \\ &= s_n - (k_1 z_1 + k_2 z_2 + \dots + k_n z_n) \end{aligned} \quad (4.8.2)$$

The mean-square error is then

$$\begin{aligned} E(e_n^2) &= E[s_n - (k_1 z_1 + k_2 z_2 + \dots + k_n z_n)]^2 \\ &= E(s_n^2) + [k_1^2 E(z_1^2) + k_2^2 E(z_2^2) + \dots + k_n^2 E(z_n^2) \\ &\quad + 2k_1 k_2 E(z_1 z_2) + 2k_1 k_3 E(z_1 z_3) + \dots] \\ &\quad - [2k_1 E(z_1 s_n) + 2k_2 E(z_2 s_n) + \dots + 2k_n E(z_n s_n)] \end{aligned} \quad (4.8.3)$$

We now wish to find k_1, k_2, \dots, k_n such as to minimize $E(e_n^2)$. The usual methods of differential calculus lead to the following set of linear equations:

$$\begin{bmatrix} E(z_1^2) & E(z_1 z_2) & \dots & \dots \\ E(z_2 z_1) & E(z_2^2) & & \\ \vdots & \vdots & & \\ E(z_n z_1) & E(z_n z_2) & \dots & \dots \\ & E(z_n^2) & & \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix} = \begin{bmatrix} E(z_1 s_n) \\ E(z_2 s_n) \\ \vdots \\ E(z_n s_n) \end{bmatrix} \quad (4.8.4)$$

Just as in the continuous problem, we assume that the auto- and cross-correlation functions of the signal and noise are known, so that all the expectations indicated in Eq. (4.8.4) are available and the equations may be solved for the weighting factors k_1, k_2, \dots, k_n . Thus, in principle at least, we have a formal solution. The indicated solution may even be practical in off-line situations where the amount of measurement data is limited. However, in on-line situations, the solution keeps growing in size as time evolves. With each new measurement, we have to add an additional equation, compute a new row of covariances, and re-solve the resultant equations for a new set of k 's. The solution gets out of hand in a hurry! Even with clever programming in the solution of Eq. (4.8.4), one cannot avoid the "growing memory" problem caused by the need to store all of the past measurement data. One has no right, a priori at least, to discard any of the past input data, no matter how old it may be. There are many situations where "old" data is just as good as "new" data. The net result is that the weighting function approach is, at best, unwieldy in on-line applications. We will see in Chapter 5 that the discrete Kalman filter provides a solution that is readily implemented on a computer, and usually it is preferred over the direct weighting-function solution given by Eq. (4.8.4).

4.9 PERSPECTIVE

A Wiener filter minimizes the mean-square estimation error subject to certain constraints and assumptions. It is important to remember that this optimization is only intended to apply to the problem of separating one *noiselike* signal from another, which is a very restricted class of filtering problems. Also the assumption of *linear* filtering was built into the derivation from the start. We will see later in Chapter 5 that this is not a serious restriction if all the random processes involved are Gaussian. In this one case the linear filter is optimum by most any reasonable criterion of performance (13). However, the non-Gaussian case is another matter. A nonlinear filter may be better, and the Wiener filter is only optimal within the restricted class of linear filters.

Sometimes, minimization of the mean-square error can lead to seemingly strange physical results; therefore, the results of Wiener filtering should be viewed with a degree of caution. An example will illustrate this. It was mentioned in Section 2.11 that both the random telegraph wave and the Gauss-Markov process have exponential autocorrelation functions. Since the Wiener filter design depends only on the auto- and cross-correlation functions, the solution for the pure prediction problem will be the same for both random-telegraph and Gauss-Markov signals with the same autocorrelation functions. It was found in Example 4.4 that the solution in this case is a simple attenuator and, for large prediction time, the predictor output is approximately zero. This makes good sense for the Gauss-Markov signal, because it is noiselike with a central tendency toward zero; and, in the absence of relevant (i.e., "recent") measurement information, one should just pick the process mean as the best estimate. This way the estimate is only rarely in error by a gross amount, and it is often close to the signal value. On the other hand, to estimate the random telegraph signal to be zero is pure nonsense. We know a priori that the signal is either +1 or -1, and it is *never* zero. That is, the Wiener predictor never predicts the correct answer, nor is it even close to the correct answer! We might better arbitrarily choose +1 as our estimate. Then, at least we would be correct half the time! Think of the many game (and more serious) situations where you would be better off to be exactly correct half the time (and grossly in error the other half) than to be significantly (but perhaps not grossly) in error *all* the time.

The reason for the strange result in the random-telegraph-wave predictor is that the optimization procedure did not account for the higher-order "statistics" of the process. The mean-square-error criterion is simplistic and only calls for knowledge of the correlation functions of the processes involved. This is, of course, convenient and it also happens to fit the Gaussian process case quite well (for reasons that may not be apparent yet). However, as has just been demonstrated, the mean-square-error criterion of performance can lead to strange results when dealing with non-Gaussian processes.

Problems

4.1. The first-order filter assumed in Example 4.1 will be recognized as a first-order Butterworth filter. Variation of the time constant T in the optimization process is the equivalent of varying the -3-db cutoff frequency of the filter. It was noted later in Example 4.2 that for $\sigma^2 = \beta = A = 1$, the Phillips procedure led to a cutoff frequency of $1/(1 + \sqrt{2})$ rad/sec and a mean-square error of .914.

In this problem, consider a second-order Butterworth filter of the form

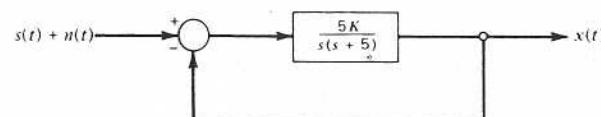
$$G(s) = \frac{1}{\left(\frac{s}{\omega_c}\right)^2 + \sqrt{2}\left(\frac{s}{\omega_c}\right) + 1} \quad (\text{P4.1})$$

Just as in the first-order case, only one parameter is free to vary—the -3-db cutoff frequency ω_c . Using the same noise parameters as before (i.e., $\sigma^2 = \beta = A = 1$), find the optimum cutoff frequency for the second-order Butterworth filter. Also find the associated mean-square error and compare the results with those obtained for the first-order filter. (Note: Even a second-order filter with only one adjustable parameter involves a considerable amount of algebraic effort. It might be easier to find the minimum by trial-and-error, numerical means rather than by analytical methods. Also note that the resulting second-order filter may not be better than the first-order filter because of the constraints placed on the form of the filter.)

4.2. A closed-loop position control system has the form shown in the block diagram. The spectral density function of the derivative of the signal $s(t)$ is given by

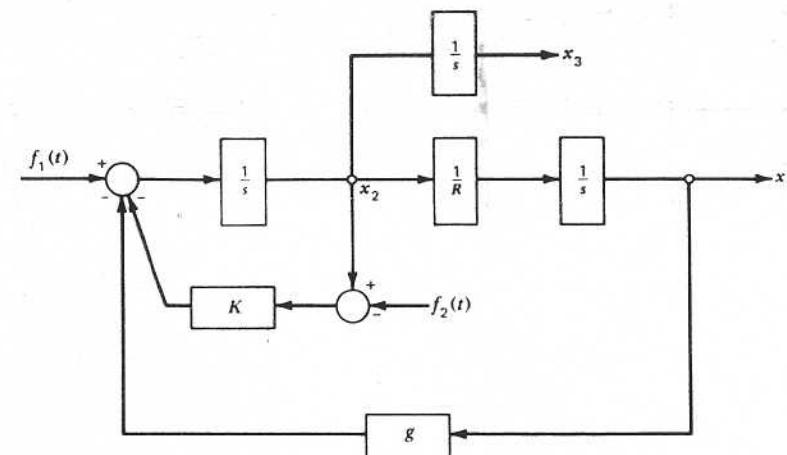
$$S_s(j\omega) = \frac{1000}{\omega^2 + .01}$$

and the noise is unity-amplitude Gaussian white noise. Find the value of the gain constant K that will minimize the mean-square error, and find the damping ratio corresponding to this gain. (Hint: The error term that involves the signal spectrum is of the form $S(s)[1 - G(s)]$ [see Eq. (4.2.3)]. In this problem $[1 - G(s)]$ contains an s factor in the numerator that may be linked with $S(s)$. Since this has the interpretation of derivative of $s(t)$ in the time domain, the mean-square error term due to signal may be written in terms of the spectrum of the derivative of the signal, which is the spectral function given in the problem. This problem is worked out as an example in J. G. Truxal, *Automatic Feedback Control System Synthesis*, New York, McGraw-Hill, 1955, pp. 472-477.)



Problem 4.2

4.3. The figure for Problem 3.21 is repeated here for convenience. Let $f_1(t)$ and $f_2(t)$ be independent white noise inputs with spectral amplitudes A_1 and A_2 , just as in Problem 3.21. Consider x_3 (the inertial system position error) as the output and find the value of K that minimizes the mean-square value of x_3 .



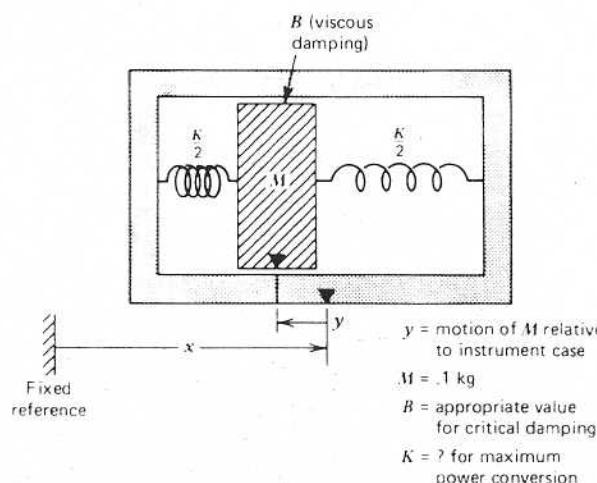
R = earth radius $\approx 2.09 \times 10^7$ ft

g = gravitational constant
 ≈ 32.2 ft/sec²

K = feedback constant
(adjustable to yield the desired damping ratio)

Problem 4.3

4.4. One facet of biomedical electronic instrumentation work is that of implanting miniature, telemetering transducers in live animals. This is done in order that various body functions may be observed under normal conditions of activity and environment. When considering such an implant, one is immediately confronted with the problem of



Problem 4.4

supplying energy to the transducer. Batteries are an obvious possibility, but they are bulky and have finite life. Another possibility is to take advantage of the random motion of the animal, and a device for accomplishing this is shown in the simplified diagram of the accompanying figure. The energy conversion takes place in the damping mechanism. This is shown as simple viscous damping, but, in fact, would have to be some sort of electromechanical conversion device. Assuming that all the power indicated as being dissipated in damping can be converted to electrical form, what is the optimum value for the spring constant K and how much power is converted for this optimum condition? The spring-mass arrangement is to be critically damped, the mass is .1 kg, and the autocorrelation function for the velocity \dot{x} is estimated to be

$$R_s(\tau) = 1e^{-2\pi|\tau|} \text{ (ft/sec)}^2$$

[For more details see "Biological Power Sources," F. M. Long, Ph.D. Thesis (Supervised by V. W. Bolie), Iowa State University, 1961.]

- 4.5. Find the optimal causal transfer function for the case where the autocorrelation functions of the signal and noise are

$$R_s(\tau) = 2e^{-2|\tau|}$$

$$R_n(\tau) = e^{-|\tau|}$$

Also find the mean-square error. The signal $s(t)$ and the noise $n(t)$ may be assumed to be independent of each other and both are time stationary. Also let the prediction time be zero. In brief, this is the classical Wiener filter problem with zero prediction time.

- 4.6. Find the optimal noncausal $G(s)$ for the signal and noise situation given in Problem 4.5. Also, find the corresponding weighting function and the resultant mean-square error.

- 4.7. Consider a stationary Gaussian signal whose spectral density function is

$$S_s(j\omega) = \frac{\omega^2 + 1}{\omega^4 + 8\omega^2 + 16}$$

Find the optimal predictor for the stationary case and $\alpha = 1$. (Note: The causal solution is desired, and the predictor may be specified either in terms of a transfer function or weighting function.)

- 4.8. Consider an additive combination of signal and noise where the spectral densities are given by

$$S_s(s) = \frac{2}{-s^2 + 1}$$

$$S_n(s) = \frac{4}{-s^2 + 4}$$

Both the signal and noise are stationary Gaussian processes, and they are statistically independent.

- (a) Without regard to causality, what is the optimal linear filter for the stationary case? The answer may be specified either in terms of a transfer function or weighting function.
- (b) Does the noncausal result of (a) have any physical significance? That is, is there a corresponding estimation problem where the computed theoretical mean-square

error would have the significance of actual mean-square estimation error? Explain briefly.

- 4.9. Consider the following autocorrelation functions of the signal and noise in a stationary Wiener prediction problem

$$R_s(\tau) = 4e^{-4|\tau|}$$

$$R_n(\tau) = e^{-|\tau|}$$

The signal and noise are independent and the prediction time is .25 sec. Find the optimal causal weighting function.

- 4.10. Show that the Wiener-Hopf equation [Eq. (4.3.18)] is a sufficient as well as a necessary condition for minimizing the mean-square error. [Hint: Differentiate Eq. (4.3.6) twice with respect to ϵ and then examine $d^2E(\epsilon^2)/d\epsilon^2$. Recall from elementary calculus that the extremum is a relative minimum if this quantity is positive at the extremum point, which, in this case is $\epsilon = 0$.]

- 4.11. The orthogonality principle states that the filter error at the current time t is orthogonal to the filter input evaluated at any previous time since initiation of the input. In Section 4.5 this principle was derived from the nonstationary version of the Wiener-Hopf equation [Eq. (4.4.4)]. Reverse the arguments and show that the integral equation [i.e., Eq. (4.4.4)] may be derived from the orthogonality principle. (If you need help on this one, see W. B. Davenport, Jr., and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958, p. 240.)

- 4.12. Consider a noisy measurement of the form

$$z(t) = a_0 + n(t)$$

where a_0 is an unknown random constant with zero-mean normal distribution and a variance of σ^2 . The additive noise may be assumed to be white with spectral amplitude A . Find the optimal time-variable filter for estimating a_0 . The filter is turned on at $t = 0$. (Hint: A random constant may be thought of as a limiting case of a Markov process with a very large time constant.)

- 4.13. Verify that the $a(t)$ and $b(t)$ coefficients given in Example 4.5 are correct.

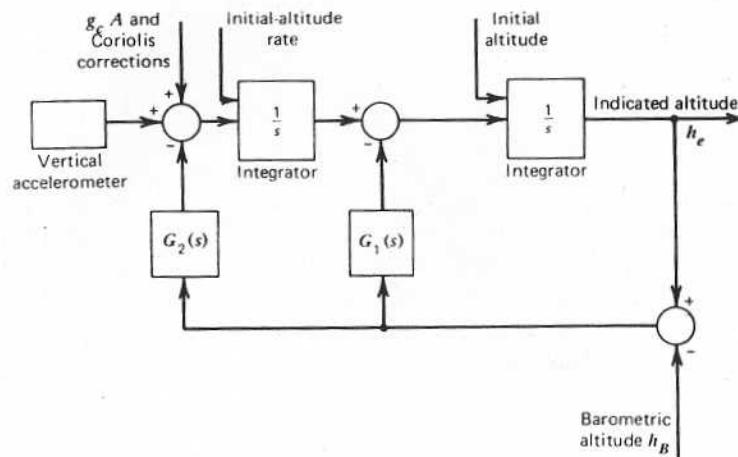
- 4.14. The accompanying figure was taken from M. Kayton and W. R. Fried (Eds.), *Avionics Navigation Systems*, New York: Wiley, 1969, p. 318. It describes a means of blending together barometrically derived and inertially derived altitude signals in such a way as to take advantage of the best properties of both signals. The barometric signal, when taken by itself, has large inherent time lag that is undesirable. It is, however, stable and reasonably accurate in the steady-state. On the other hand, integrated vertical acceleration has an unbounded steady-state error because of accelerometer bias error. In effect, the high-frequency response of the accelerometer is good, but its low-frequency response is poor; just the reverse is true for the barometric instrument. Thus this is an ideal setting for a complementary filter application.

Let $G_1(s)$ and $G_2(s)$ in the figure be constant gains G_1 and G_2 and show that:

- (a) The system fits the form of a complementary filter as discussed in Section 4.6.
- (b) The system error has second-order characteristics with the natural frequency and damping ratio being controlled by the designer's choice of G_1 and G_2 .

[Note: In order to show the complementary-filter property, you must conceptually think of directly integrating the accelerometer signal twice ahead of the summer to obtain an inertially derived altitude signal contaminated with error. However, direct

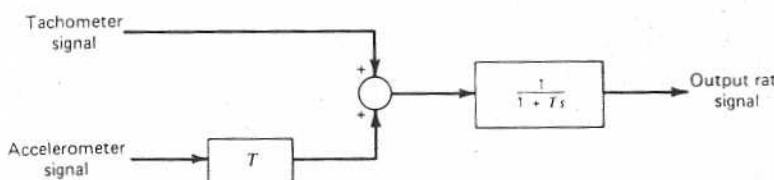
integration of the accelerometer output is not required in the final implementation because of cancellation of s 's in numerator and denominator of the transfer function. This will be apparent after carrying through the details of the problem. Also note that the g_c , A , and Coriolis corrections indicated in the figure are simply the gravity, accelerometer-bias, and Coriolis corrections required in order that the accelerometer output be \bar{h} (as best possible). Also, the initial conditions indicated in the figure may be ignored because the system is stable and we are only interested in the steady-state condition here.



Problem 4.14

4.15. It is desired to add rate feedback in a control system. The rate signal is to come from a permanent-magnet, dc tachometer on the output shaft. However, it is noticed that the tachometer output is noisy due to the combined commutator-brush action. Filtering the tachometer output is a possibility, but this introduces unwanted delay in the rate signal. Someone suggests that if we add an angular accelerometer as well as the dc tachometer on the output shaft, we can filter the rate signal without paying the price of time delay. The proposed scheme is shown in the accompanying figure.

- Show that this is a form of complementary filtering.
- Assume that the errors associated with the tachometer and accelerometer signals are independent white noise processes with spectral amplitudes K_T and K_A , respectively. Find the value of T that minimizes the mean-square error.



3. R. E. Kalman and R. Bucy, "New Results in Linear Filtering and Prediction," *Trans. of the ASME—Journal of Basic Engineering*, 83, 95 (1961).
4. H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of Servomechanisms* (Vol. 25), Radiation Laboratory Series, New York: McGraw-Hill, 1947.
5. H. W. Bode and C. E. Shannon, "A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory," *Proc. I.R.E.*, 38, 417–424 (April 1950).
6. G. R. Cooper and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, New York: Holt, Rinehart, and Winston, 1971.
7. R. Weinstock, *Calculus of Variations*, New York: McGraw-Hill, 1952.
8. W. B. Davenport, Jr., and W. L. Root, *Introduction to Random Signals and Noise*, New York: McGraw-Hill, 1958.
9. C. T. Chen, *Introduction to Linear System Theory*, New York: Holt, Rinehart, and Winston, 1970.
10. W. H. Wirkler, "Aircraft Course Stabilization Means," U.S. Patent 2,548,278, issued April 10, 1951.
11. W. G. Anderson and E. H. Fritze, "Instrument Approach System Steering Computer," *Proc. of I.R.E.*, 41: (February 1953).
12. R. G. Brown and J. W. Nilsson, *Introduction to Linear Systems Analysis*, New York: Wiley, 1962.
13. J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*, New York: McGraw-Hill, 1969.

CHAPTER 5

The Discrete Kalman Filter

The end result of the Wiener solution of the optimal filter problem is a filter weighting function.* In effect this tells how the past values of the input should be weighted in order to determine the present value of the output, that is, the optimal estimate. Unfortunately, the Wiener solution does not lend itself very well to the corresponding discrete-data problem nor is it easily extended to more complicated multiple-input multiple-output problems. In 1960, R. E. Kalman provided an alternative way of formulating the least squares filtering problem using state-space methods (1). Engineers, especially in the field of navigation, were quick to see the Kalman technique as a practical solution to a number of problems that were previously considered intractable using Wiener methods. The two main features of the Kalman formulation and solution of the problem are (1) vector modeling of the random processes under consideration and (2) recursive processing of the noisy measurement (input) data. We now proceed to the details of the recursive solution of the discrete-data linear filter problem.

5.1 A SIMPLE RECURSIVE EXAMPLE

When working with practical problems involving discrete data, it is important that our methods be computationally feasible as well as mathematically correct. A simple example will illustrate this. Consider the problem of estimating the mean of some unknown constant based on a sequence of noisy measurements. Let us assume that our estimate is to be the sample mean and that we wish to refine our estimate with each new measurement as it becomes available. That is, think of processing the data on-line. Let the measurement sequence be denoted as z_1, z_2, \dots, z_n , where the subscript denotes the time at which the measurement is taken. One method of processing the data would be to store each measurement as it becomes available and then compute the sample mean in accordance with the following algorithm (in words):

1. First measurement z_1 : Store z_1 and estimate the mean as

$$\hat{m}_1 = z_1$$

*Even though the details of Wiener filter theory are not needed for this and subsequent chapters, the reader should at least browse through Sections 4.1 and 4.2 in order to understand the problem statement and assumptions used in the Wiener solution.

2. Second measurement z_2 : Store z_2 along with z_1 and estimate the mean as

$$\hat{m}_2 = \frac{z_1 + z_2}{2}$$

3. Third measurement z_3 : Store z_3 along with z_1 and z_2 and estimate the mean as

$$\hat{m}_3 = \frac{z_1 + z_2 + z_3}{3}$$

4. And so forth.

Clearly, this would yield the correct sequence of sample means as the experiment progresses. It should also be clear that the amount of memory needed to store the measurements keeps increasing with time, and also the number of arithmetic operations needed to form the estimate increases correspondingly. This would lead to obvious problems when the total amount of data is large. Thus, consider a simple variation in the computational procedure in which each new estimate is formed as a blend of the old estimate and the current measurement. To be specific, consider the following algorithm:

1. First measurement z_1 : Compute the estimate as

$$\hat{m}_1 = z_1$$

Store \hat{m}_1 and discard z_1 .

2. Second measurement z_2 : Compute the estimate as a weighted sum of the previous estimate \hat{m}_1 and the current measurement z_2 :

$$\hat{m}_2 = \frac{1}{2} \hat{m}_1 + \frac{1}{2} z_2$$

Store \hat{m}_2 and discard z_2 and \hat{m}_1 .

3. Third measurement z_3 : Compute estimate as a weighted sum of \hat{m}_2 and z_3 :

$$\hat{m}_3 = \frac{2}{3} \hat{m}_2 + \frac{1}{3} z_3$$

Store \hat{m}_3 and discard z_3 and \hat{m}_2 .

4. And so forth. It should be obvious that at the n th stage the weighted sum is

$$\hat{m}_n = \left(\frac{n-1}{n} \right) \hat{m}_{n-1} + \left(\frac{1}{n} \right) z_n$$

Clearly, the above procedure yields the same identical sequence of estimates as before, but without the need to store all the previous measurements. We simply use the result of the previous step to help obtain the estimate at the current step of the process. In this way, the previous computational effort is used to good advantage and not wasted. The second algorithm can proceed on ad infinitum without a growing memory problem. Eventually, of course, as n becomes extremely large, a round-off problem might be encountered. However, this is to be expected with either of the two algorithms.

The second algorithm is a simple example of a *recursive* mode of operation. The key element in any recursive procedure is the use of the results of the previous step to aid in obtaining the desired result for the current step. This is one of the main features of Kalman filtering, and one that clearly distinguishes it from the weighting-function (Wiener) approach, which requires arithmetic operations on all the past data.

In order to apply the recursive philosophy to estimation of a random process, it is first necessary that both the process and the measurement noise be modeled in vector form. Thus we digress for a moment and look at the vector description of a random process. We will then return to the recursive filtering problem in Section 5.4.

5.2 VECTOR DESCRIPTION OF A CONTINUOUS RANDOM PROCESS

In the material to follow, it is desirable to have the random process description written in vector form. We begin with a scalar process $y(t)$ for which we have a description as discussed in Chapter 2. We now wish to rewrite the description in the following format:

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{w} \quad (5.2.1)$$

$$\mathbf{y} = \mathbf{B}\mathbf{x} \quad (5.2.2)$$

The vector process under consideration is denoted as \mathbf{x} and is an $(n \times 1)$ column vector. \mathbf{F} , \mathbf{G} , and \mathbf{B} are rectangular matrices whose elements may be time varying, and \mathbf{w} is a $(p \times 1)$ column vector whose elements are white noise processes. This is called the *state model* of the process, and, in words, it simply says that we want to think of our process as the result of passing white noise through some system with linear dynamics.* The original process $y(t)$ is required to be a linear combination of the system state variables via Eq. (5.2.2) and, thus, the old process can be recovered from the new model with the appropriate additive combination of state variables.

In general it is not possible to model all random processes in the form of Eqs. (5.2.1) and (5.2.2). However, many of the processes encountered in physical applications may be so modeled; and, in particular, all processes with rational spectral density functions have state models with finite dimensionality. A procedure for forming the state model for such processes follows.

*Only an elementary knowledge of state-space methods is needed for our discussion of Kalman filtering. The treatment of state-space methods given in most introductory books on linear control theory (e.g., Refs. 2 and 3) is quite adequate for our purposes here.

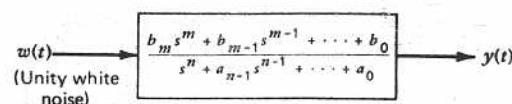


Figure 5.1. Shaping filter ($m = n - 1$).

State Model for a Process with a Rational Spectral Function

Assume we know the spectral density function $S_y(s)$ for the process $y(t)$. We first factor the spectral function using the same spectral factorization methods used in Chapters 3 and 4. This leads to

$$S_y(s) = S_y^+(s) S_y^-(s) \quad (5.2.3)$$

where we are assured that both the poles and zeros of S_y^+ lie in the left half plane. Now, it is clear from Section 3.6 that $S_y^+(s)$ is just the shaping filter required to shape unity white noise into the process $y(t)$. The assumed form of the shaping-filter transfer function [i.e., $S_y^+(s)$] is shown in Fig. 5.1, and it is assumed that the order of the denominator is at least one greater than the numerator. This is necessary in order for the $y(t)$ process to have finite variance.

It is now convenient to decompose the block diagram of Fig. 5.1 into the equivalent form shown in Fig. 5.2. With the state variables chosen as the usual phase variables, we are assured that each will have bounded variance. The final state model for the process $y(t)$ is then

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & & & \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} w(t) \quad (5.2.4)$$

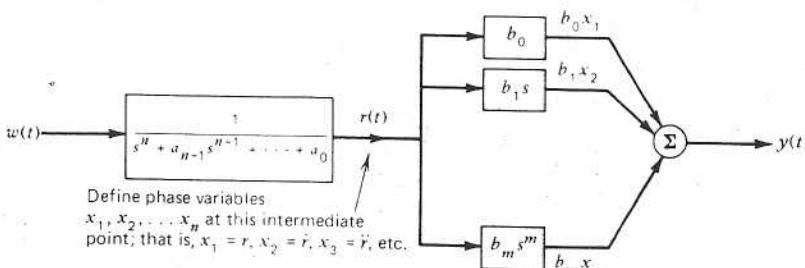


Figure 5.2. Block diagram showing state variables.

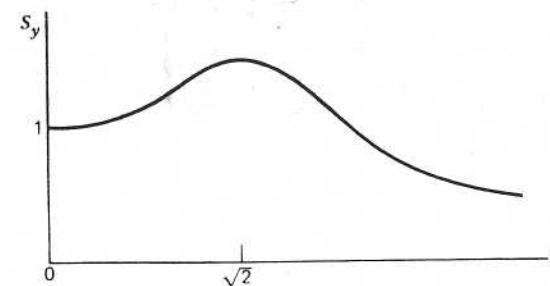


Figure 5.3. Spectral function for Example 5.1.

$$y = [b_0 \ b_1 \ \dots \ b_{n-1}] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (5.2.5)$$

Two examples illustrate this modeling procedure.

Example 5.1 Consider a random process with a spectral function given by

$$S_y(j\omega) = \frac{16}{\omega^4 - 4\omega^2 + 16} \quad (5.2.6)$$

A sketch of this spectral function is shown in Fig. 5.3 and it shows a mild resonance phenomenon occurring at $\sqrt{2}$ rad/sec. We first need to convert $S_y(j\omega)$ to the s domain by replacing $j\omega$ with s . The result is

$$S_y(s) = \frac{16}{s^4 + 4s^2 + 16} \quad (5.2.7)$$

Since the poles of $S_y(s)$ are at $s = -1 \pm j\sqrt{3}$ and $s = 1 \pm j\sqrt{3}$, $S_y(s)$ can be factored in the form

$$S_y(s) = S_y^+(s) \cdot S_y^-(s) = \frac{4}{s^2 + 2s + 4} \cdot \frac{4}{(-s)^2 + 2(-s) + 4} \quad (5.2.8)$$

A block diagram showing the state and output relationships can now be formed as shown in Fig. 5.4.

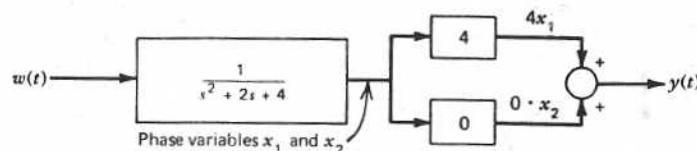


Figure 5.4. Block diagram for state model for Example 5.1.

The complete state model is then

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w(t) \quad (5.2.9)$$

$$y = [4 \quad 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5.2.10)$$

■

Example 5.2 The Gauss-Markov process model is frequently used in engineering applications. This is partly because of its simplicity and partly because the analyst often lacks the detailed knowledge of the process under consideration that would be needed to devise a more complicated model. The Gauss-Markov process has an autocorrelation function given by

$$R_y(\tau) = \sigma^2 e^{-\beta|\tau|}$$

The corresponding spectral function is

$$S_y(s) = \frac{2\sigma^2\beta}{-s^2 + \beta^2} = \frac{\sqrt{2\sigma^2\beta}}{s + \beta} \frac{\sqrt{2\sigma^2\beta}}{-s + \beta}$$

Clearly, the shaping filter is

$$S_y^+(s) = \frac{\sqrt{2\sigma^2\beta}}{s + \beta}$$

and the corresponding differential equation for y is

$$\dot{y} = -\beta y + \sqrt{2\sigma^2\beta} w(t) \quad (5.2.11)$$

where $w(t)$ is unity white noise. Equation (5.2.11) is first order, so this is the state model for the process. ■

Nonstationary and Deterministic Processes

The state model of a random process is reasonably general in form and will accommodate a wide variety of situations. All that is required is that the pro-

cess under consideration be related to white noise via a linear differential equation. A few examples of processes that do not have rational spectral functions will illustrate this further.

1. Wiener process (Brownian motion): The Wiener process is defined as the integral of Gaussian white noise with zero initial condition. Thus, the appropriate state model is first order and is

$$\dot{y} = w(t) \quad y(0) = 0 \quad (5.2.12)$$

where $w(t)$ is Gaussian white noise with known spectral amplitude.

2. Random bias: A random constant satisfies the differential equation

$$\dot{y} = 0 \quad y(0) = a_0 \quad (5.2.13)$$

The initial condition a_0 is a random variable whose distribution is presumed to be known.

3. Random ramp: A random ramp process with random initial value and slope may be written as

$$y(t) = a_0 + a_1 t \quad (5.2.14)$$

where a_0 and a_1 are random variables. The differential equation corresponding to Eq. (5.2.14) is

$$\ddot{y} = 0 \quad y(0) = a_0 \quad \dot{y}(0) = a_1 \quad (5.2.15)$$

This is a second-order differential equation, so the state vector for the process y must be a two-tuple. Using phase variables in the vector model leads to

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} w, \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} \quad (5.2.16)$$

$$y = [1 \quad 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5.2.17)$$

4. Harmonic motion with random amplitude and phase: A process having harmonic motion with known frequency ω_0 rad/sec satisfies the differential equation

$$\ddot{y} + \omega_0^2 y = 0 \quad (5.2.18)$$

The solution of this equation is

$$y(t) = y(0) \cos \omega_0 t + \frac{\dot{y}(0)}{\omega_0} \sin \omega_0 t \quad (5.2.19)$$

To specify "random amplitude and phase" is equivalent to saying "random sine and cosine components." Thus, choosing phase variables as the state

variables in the vector model, we have

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} w, \quad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} \quad (5.2.20)$$

$$y = [1 \ 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5.2.21)$$

The random variables A and B are assigned probability density functions in accordance with the given distributions of the sine and cosine components of y . A common assumption is to let A and B be independent zero-mean normal random variables with variances σ^2 and $\omega_0^2\sigma^2$. The resultant harmonic motion then has Rayleigh amplitude and uniform phase distributions.

The preceding examples are just a few of the many processes that can be modeled in vector form. Stationarity is not necessary. The only requirement is that the process must somehow-or-other be related to white noise through a linear differential equation. It is especially important to note that one does not choose the number of state variables in the model at will; rather, the number of elements in the state vector is fixed by the order of the differential equations relating the various processes in the model to the white-noise driving functions. This is a basic fact of life and must be adhered to if the state model is to faithfully represent the situation at hand. Engineering literature abounds with examples of sloppy modeling and the reader is cautioned accordingly.

Also, remember that the vector model of a random process is not unique. You can always perform any nonsingular linear transformation on a set of state variables and obtain another valid set. Of course, since the F , G , and B matrices [see Eqs. (5.2.1) and (5.2.2)] for the transformed set will be different from those of the original set, the new model will look different. It will, however, be a perfectly proper model if the transformation is done correctly. Problems 5.7, 5.8, and 5.9 elaborate on this further.

5.3 DISCRETE-TIME MODEL

Discrete-time processes may arise in either of two ways. First, there is the situation where a sequence of events takes place naturally in discrete steps. This might be the result of a sequence of chance experiments such as the discrete random-walk problems of statistics. Recall in this problem, we imagine the walker taking a number of steps at random, either forward or backward. The length of each step may be either a fixed or a random variable. In either case, the random variable of interest is the distance from the origin after taking n steps. In this problem, there is no such thing as fractional steps. The time variable moves in discrete jumps!

Discrete-time processes may also arise from sampling a continuous pro-

cess at discrete times. The sampling may be intentional and may be under the control of the designer, as is the case when analog data is converted to digital form. Or, sometimes the sampling may be unintentional and forced on us by a measurement constraint that only allows observation of the process at discrete points in time. For example, the TRANSIT satellite navigation system provides the user with only a single position fix with each satellite pass (4). Thus the user is only allowed to observe his or her position at discrete points in time. Furthermore, in this case the observation times are not equally spaced because the various satellite orbits do not have perfect symmetry.

Irrespective of how the discretization arises in the physical problem, we wish to fit all situations into the following format:

$$\mathbf{x}_{k+1} = \phi_k \mathbf{x}_k + \mathbf{w}_k \quad (5.3.1)$$

$$\mathbf{y}_k = \mathbf{B}_k \mathbf{x}_k \quad (5.3.2)$$

where

\mathbf{x}_k = vector state of the process at time t_k

ϕ_k = matrix that relates \mathbf{x}_k to \mathbf{x}_{k+1} in the absence of a forcing function. (In the sampled version of a continuous process this is the state transition matrix)

\mathbf{w}_k = timewise uncorrelated zero-mean sequence (i.e., a discrete process with zero autocorrelation function, except for $\tau = 0$)

Just as in the continuous model, \mathbf{B}_k provides the linear connection between the system state vector and the scalar process being modeled. Note that we allow correlation among the elements of \mathbf{w}_k at any point in time t_k , but we demand that the sequence be uncorrelated timewise. We assume that the covariance structure of \mathbf{w}_k is known, and it will be denoted as \mathbf{Q}_k . Thus we have

$$E[\mathbf{w}_k \mathbf{w}_i^T] = \begin{cases} \mathbf{Q}_k, & i = k \\ 0, & i \neq k \end{cases} \quad (5.3.3)$$

where "super T " denotes transpose. (We will reserve "prime" for other uses.)

As mentioned previously, the discrete model of Eqs. (5.3.1) and (5.3.2) need not come from a sampled continuous situation, but this is so common in applied work that it warrants amplification. Let us say we begin with a continuous process described by

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{w} \quad (5.3.4)$$

where \mathbf{w} is a vector forcing function whose elements are white noise. Let us consider samples of this process at discrete times $t_0, t_1, \dots, t_k, \dots$. State-space methods may be used to obtain the difference equation relating the samples of \mathbf{x} . Specifically, the solution of Eq. (5.3.4) at time t_{k+1} may be written as

$$\mathbf{x}(t_{k+1}) = \phi(t_{k+1}, t_k)\mathbf{x}(t_k) + \int_{t_k}^{t_{k+1}} \phi(t_{k+1}, \tau)\mathbf{G}(\tau)\mathbf{w}(\tau) d\tau \quad (5.3.5)$$

Or, in our abbreviated notation,

$$\mathbf{x}_{k+1} = \phi_k \mathbf{x}_k + \mathbf{w}_k$$

Clearly, ϕ_k is the state transition matrix for the step from t_k to t_{k+1} , and \mathbf{w}_k is the driven response at t_{k+1} due to the presence of the white noise input during the (t_k, t_{k+1}) interval. Note that the white-noise input requirement in the continuous model automatically assures that \mathbf{w}_k will be an uncorrelated (white) sequence in the discrete model.

Analytical methods for finding the state transition matrix are well known and these may be used in systems with low dimensionality. However, evaluation of the \mathbf{Q}_k matrix that describes \mathbf{w}_k may not be so obvious. Formally, we can write \mathbf{Q}_k in integral form as

$$\begin{aligned}\mathbf{Q}_k &= E[\mathbf{w}_k \mathbf{w}_k^T] \\ &= E\left\{\left[\int_{t_k}^{t_{k+1}} \phi(t_{k+1}, u) \mathbf{G}(u) \mathbf{w}(u) du\right] \left[\int_{t_k}^{t_{k+1}} \phi(t_{k+1}, v) \mathbf{G}(v) \mathbf{w}(v) dv\right]^T\right\} \\ &= \int_{t_k}^{t_{k+1}} \int_{t_k}^{t_{k+1}} \phi(t_{k+1}, u) \mathbf{G}(u) E[\mathbf{w}(u) \mathbf{w}^T(v)] \mathbf{G}^T(v) \phi^T(t_{k+1}, v) du dv \quad (5.3.6)\end{aligned}$$

The matrix $E[\mathbf{w}(u) \mathbf{w}^T(v)]$ is a matrix of Dirac delta functions that, presumably, is known from the continuous model. Thus, in principle, \mathbf{Q}_k may be evaluated from Eq. (5.3.6). This is not a trivial task, though, even for low-order systems. If the continuous system giving rise to the discrete situation has constant parameters and if the various white noise inputs are uncorrelated, some simplification is possible and the transfer function methods of Chapter 3 may be applied. This is best illustrated with an example rather than in general terms.

Example 5.3 The integrated Gauss-Markov process shown in Fig. 5.5 is frequently encountered in engineering applications. The continuous model in this case is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -\beta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \sqrt{2\sigma^2\beta} \end{bmatrix} w(t) \quad (5.3.7)$$

$$y = [1 \ 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5.3.8)$$

Let us say the sampling interval is Δt and we wish to find the corresponding discrete model. This amounts to the determination of ϕ_k , \mathbf{Q}_k , and \mathbf{B}_k . The

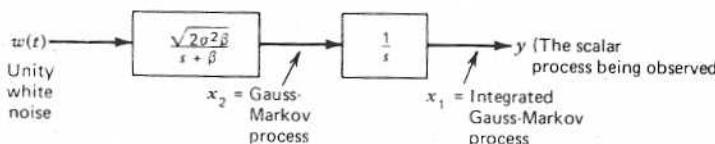


Figure 5.5. Integrated Gauss-Markov process.

transition matrix is easily determined as

$$\begin{aligned}\phi_k &= [\mathcal{L}^{-1} [(s\mathbf{I} - \mathbf{F})^{-1}]]_{t=\Delta t} \\ &= \mathcal{L}^{-1} \begin{bmatrix} s & -1 \\ 0 & s + \beta \end{bmatrix}^{-1} = \mathcal{L}^{-1} \begin{bmatrix} \frac{1}{s} & \frac{1}{s(s + \beta)} \\ 0 & \frac{1}{s + \beta} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{1}{\beta} (1 - e^{-\beta\Delta t}) \\ 0 & e^{-\beta\Delta t} \end{bmatrix} \quad (5.3.9)\end{aligned}$$

Next, rather than using Eq. (5.3.6) directly to determine \mathbf{Q}_k , we use the transfer function approach. From the block diagram of Fig. 5.5 we observe the following transfer functions:

$$G(w \text{ to } x_1) = G_1 = \frac{\sqrt{2\sigma^2\beta}}{s(s + \beta)} \quad (5.3.10)$$

$$G(w \text{ to } x_2) = G_2 = \frac{\sqrt{2\sigma^2\beta}}{s + \beta} \quad (5.3.11)$$

The corresponding weighting functions are

$$g_1(t) = \sqrt{\frac{2\sigma^2}{\beta}} (1 - e^{-\beta t}) \quad (5.3.12)$$

$$g_2(t) = \sqrt{2\sigma^2\beta} e^{-\beta t} \quad (5.3.13)$$

We can now use the methods of Chapter 3 to find the needed mean-square responses

$$\begin{aligned}E[x_1 x_1] &= \int_0^{\Delta t} \int_0^{\Delta t} g_1(u) g_1(v) E[w(u) w(v)] du dv \\ &= \int_0^{\Delta t} \int_0^{\Delta t} \frac{2\sigma^2}{\beta} (1 - e^{-\beta u})(1 - e^{-\beta v}) \delta(u - v) du dv \\ &= \frac{2\sigma^2}{\beta} [\Delta t - \frac{2}{\beta} (1 - e^{-\beta\Delta t}) + \frac{1}{2\beta} (1 - e^{-2\beta\Delta t})] \quad (5.3.14)\end{aligned}$$

$$\begin{aligned}E[x_1 x_2] &= \int_0^{\Delta t} \int_0^{\Delta t} g_1(u) g_2(v) E[w(u) w(v)] du dv \\ &= \int_0^{\Delta t} \int_0^{\Delta t} 2\sigma^2 e^{-\beta u} (1 - e^{-\beta v}) \delta(u - v) du dv \\ &= 2\sigma^2 \left[\frac{1}{\beta} (1 - e^{-\beta\Delta t}) - \frac{1}{2\beta} (1 - e^{-2\beta\Delta t}) \right] \quad (5.3.15)\end{aligned}$$

$$\begin{aligned}
 E[x_2 x_2] &= \int_0^{\Delta t} \int_0^{\Delta t} g_2(u) g_2(v) E[w(u) w(v)] du dv \\
 &= \int_0^{\Delta t} \int_0^{\Delta t} 2\sigma^2 \beta e^{-\beta u} e^{-\beta v} \delta(u - v) du dv \\
 &= \sigma^2(1 - e^{-2\beta \Delta t})
 \end{aligned} \tag{5.3.16}$$

Thus the \mathbf{Q}_k matrix is

$$\mathbf{Q}_k = \begin{bmatrix} E[x_1 x_1] & E[x_1 x_2] \\ E[x_2 x_1] & E[x_2 x_2] \end{bmatrix} = \begin{bmatrix} \text{Eq. (5.3.14)} & \text{Eq. (5.3.15)} \\ \text{Eq. (5.3.15)} & \text{Eq. (5.3.16)} \end{bmatrix} \tag{5.3.17}$$

The \mathbf{B}_k matrix is the same for both the continuous and discrete models and is

$$\mathbf{B}_k = [1 \ 0] \tag{5.3.18}$$

The discrete model is now complete with the specification of ϕ_k , \mathbf{Q}_k , and \mathbf{B}_k as given by Eqs. (5.3.9), (5.3.17), and (5.3.18). Note that the k subscript could have been dropped in this example because the sampling interval is constant. ■

Numerical Evaluation of ϕ_k and \mathbf{Q}_k

Analytical methods for finding ϕ_k and \mathbf{Q}_k work quite well for constant parameter systems with just a few elements in the state vector. However, the dimensionality does not have to be very large before it becomes virtually impossible to work out explicit expressions for ϕ_k and \mathbf{Q}_k . This is especially true if the system \mathbf{F} matrix contains time-varying terms. Thus, we often need to resort to numerical methods.

The state transition matrix (STM) tells how the dynamical system naturally evolves from one state to another in the absence of a driving function. Thus the elements of the STM are obtained as the solutions of certain initial-condition problems. For example, let the system input be zero and let the initial state at t_k be

$$\mathbf{x}(t_k) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.3.19}$$

The solution for $\mathbf{x}(t)$ for $t \geq t_k$ is then

$$\mathbf{x}(t) = \begin{bmatrix} \phi_{11}(t, t_k) & \phi_{12}(t, t_k) & \dots \\ \phi_{21}(t, t_k) & \dots & \dots \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \phi_{11}(t, t_k) \\ \phi_{21}(t, t_k) \\ \vdots \\ \vdots \end{bmatrix} \tag{5.3.20}$$

Thus the first column of the STM is obtained by solving for the system response with the initial value of \mathbf{x} set at $[1 \ 0 \ 0 \ \dots]^T$. Similarly, the second column of ϕ is obtained by letting the initial condition be $[0 \ 1 \ 0 \ \dots]^T$, and so forth. Thus, in effect, we must solve a family of initial condition problems in order to obtain all the elements of the state transition matrix. This is laborious, but fortunately numerical methods for solving the initial-condition problem of ordinary differential equations are readily available. Whole books have been written on the subject (5); therefore it will suffice to say here that the STM can usually be evaluated in a routine manner.

One special case of evaluating ϕ_k is worthy of mention at this point. If the dynamical system has fixed parameters (i.e., \mathbf{F} is a constant), the STM may be written as an exponential series.

$$\phi_k = e^{\mathbf{F}\Delta t} = \mathbf{I} + (\mathbf{F}\Delta t) + \frac{(\mathbf{F}\Delta t)^2}{2!} + \dots \tag{5.3.21}$$

where Δt is the step size. This is probably the easiest way of evaluating the STM (from the analyst's viewpoint), because Eq. (5.3.21) is easily programmed and only involves matrix-add and matrix-multiply subroutines. Equation (5.3.21) is valid for any step size Δt . However, a large number of terms in the series may be needed if Δt is large and this may lead to computational problems. Therefore, simply look at each individual case on its own merits, and then decide as to the best method of evaluating the state transition matrix.

The numerical procedure for evaluating \mathbf{Q}_k is straightforward, but laborious. In general this calls for some sort of numerical procedure for evaluating the double integral given by Eq. (5.3.6). This can be a difficult chore if the system has time-varying parameters and its dimensionality is high. An alternative approach that may be applied in some cases (and one that is easily programmed) is to subdivide the Δt interval into smaller δt steps such that only first-order terms in δt need be retained in the \mathbf{Q} expression. The contributions of each of the subintervals may then be projected ahead via a recursive equation to yield the total \mathbf{Q}_k for the whole Δt interval. An example will illustrate this procedure.

Example 5.4 Let us reconsider the integrated Gauss-Markov process used in Example 5.3. Suppose $\beta = 10 \text{ rad/sec}$ and $\Delta t = .1 \text{ sec}$. The product $\beta\Delta t$ (which is critical in this case) is certainly not small, and higher-order terms in Δt would have to be included in evaluating \mathbf{Q}_k terms such as $e^{-\beta\Delta t}$ and $e^{-2\beta\Delta t}$. Hence, in this case, we divide Δt into 1000 subintervals of equal length. The $\beta\delta t$ for each of these is .001, and it is now reasonable to neglect higher-order terms in $\beta\delta t$.

We first need to work out the "Q" for the first subinterval. Referring to Eq. (5.3.12) and Eq. (5.3.13), we note that the weighting functions g_1 and g_2 may be approximated as being zero and $\sqrt{2\sigma^2\beta}$ over the δt interval. This leads to

$$\begin{aligned} E[x_1x_1] &\approx 0 \\ E[x_1x_2] &\approx 0 \\ E[x_2x_2] &\approx (2\sigma^2\beta)\delta t \end{aligned} \quad (5.3.21)$$

The \mathbf{Q} matrix for each subinterval is then

$$\mathbf{Q}_{\text{sub}} = \begin{bmatrix} 0 & 0 \\ 0 & (2\sigma^2\beta)\delta t \end{bmatrix} \quad (5.3.22)$$

Ultimately, we want \mathbf{Q}_k for the whole interval Δt . We obtain this as the result of taking 1000 steps as follows (refer to Fig. 5.6 for notation):

Step 1 (Determination of $E[\mathbf{x}_b\mathbf{x}_b^T]$).

We have just determined that the \mathbf{Q} for a single subinterval is given by Eq. (5.3.22). Thus

$$E[\mathbf{x}_b\mathbf{x}_b^T] = \mathbf{Q}_{\text{sub}} \quad (5.3.23)$$

Step 2 (Determination of $E[\mathbf{x}_c\mathbf{x}_c^T]$).

We first note that

$$\mathbf{x}_c = \phi_{\text{sub}}\mathbf{x}_b + (\text{response during second interval}) \quad (5.3.24)$$

where ϕ_{sub} is the state transition matrix for the subinterval. Therefore,

$$E[\mathbf{x}_c\mathbf{x}_c^T] = \phi_{\text{sub}}E[\mathbf{x}_b\mathbf{x}_b^T]\phi_{\text{sub}}^T + \mathbf{Q}_{\text{sub}} \quad (5.3.25)$$

Step 3 (Determination of $E[\mathbf{x}_d\mathbf{x}_d^T]$).

Similarly, $E[\mathbf{x}_d\mathbf{x}_d^T]$ can be obtained using the result of the previous step. Thus

$$E[\mathbf{x}_d\mathbf{x}_d^T] = \phi_{\text{sub}}E[\mathbf{x}_c\mathbf{x}_c^T]\phi_{\text{sub}}^T + \mathbf{Q}_{\text{sub}} \quad (5.3.26)$$

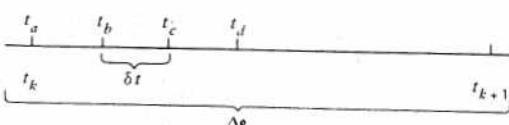


Figure 5.6. Subinterval notation for Example 5.4.

The recursive process is now clear. We simply take the previous covariance matrix, pre- and post-multiply it by ϕ_{sub} and ϕ_{sub}^T , and then add \mathbf{Q}_{sub} . After repeating this 1000 times, we have the \mathbf{Q}_k for the whole Δt interval.

If we should be in doubt about the first-order approximation in δt , we can always halve the subinterval and then carry the process through 2000 steps instead of 1000. This should lead to the same result as before. If not, we can keep halving the subinterval (and perhaps use double or triple precision) until results are consistent. This may seem to be a bit crude and wasteful of computer time. Without apology, it is. It is quite effective, though, and the programming is trivial for the analyst. In this day and age, the trade-off of computer time for analyst's time is usually a good one! ■

5.4 THE DISCRETE KALMAN FILTER

R. E. Kalman's paper describing a recursive solution of the discrete-data linear filtering problem was published in 1960 (1). About this same time, advances in digital computer technology made it possible to consider implementing his recursive solution in a number of real-time applications. This was a fortuitous circumstance, and Kalman filtering caught hold almost immediately. We will consider some examples shortly, but we must first derive the Kalman filter recursive equations which are, in effect, the "filter."

We begin by assuming the random process to be estimated can be modeled in the form

$$\mathbf{x}_{k+1} = \phi_k \mathbf{x}_k + \mathbf{w}_k \quad (5.4.1)$$

The observation (measurement) of the process is assumed to occur at discrete points in time in accordance with the linear relationship

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \quad (5.4.2)$$

Some elaboration on notation and the various terms of Eqs. (5.4.1) and (5.4.2) is in order:

\mathbf{x}_k = ($n \times 1$) process state vector at time t_k .

ϕ_k = ($n \times n$) matrix relating \mathbf{x}_k to \mathbf{x}_{k+1} in the absence of a forcing function.
(If \mathbf{x}_k is a sample of continuous process, ϕ_k is the usual state transition matrix.)

\mathbf{w}_k = ($n \times 1$) vector—assumed to be a white (uncorrelated) sequence with known covariance structure.

\mathbf{z}_k = ($m \times 1$) vector measurement at time t_k .

\mathbf{H}_k = ($m \times n$) matrix giving the ideal (noiseless) connection between the measurement and the state vector at time t_k .

$\mathbf{v}_k = (m \times 1)$ measurement error—assumed to be a white sequence with known covariance structure and uncorrelated with the \mathbf{w}_k sequence.

The covariance matrices for the \mathbf{w}_k and \mathbf{v}_k vectors are given by

$$E[\mathbf{w}_k \mathbf{w}_i^T] = \begin{cases} \mathbf{Q}_k & i = k \\ 0, & i \neq k \end{cases} \quad (5.4.3)$$

$$E[\mathbf{v}_k \mathbf{v}_i^T] = \begin{cases} \mathbf{R}_k, & i = k \\ 0, & i \neq k \end{cases} \quad (5.4.4)$$

$$E[\mathbf{w}_k \mathbf{v}_i^T] = 0, \quad \text{for all } k \text{ and } i \quad (5.4.5)$$

We assume at this point that we have an initial estimate of the process at some point in time t_k , and that this estimate is based on all of our knowledge about the process prior to t_k . This prior (*or a priori*) estimate will be denoted as $\hat{\mathbf{x}}_k^-$ where the “hat” denotes estimate, and the “super minus” is a reminder that this is our best estimate prior to assimilating the measurement at t_k . (Note that super minus as used here is not related in any way to the super minus notation used in spectral factorization.) We also assume that we know the error covariance matrix associated with $\hat{\mathbf{x}}_k^-$. That is, we define the estimation error to be

$$\mathbf{e}_k^- = \mathbf{x}_k - \hat{\mathbf{x}}_k^- \quad (5.4.6)$$

and the associated error covariance matrix is*

$$\mathbf{P}_k^- = E[\mathbf{e}_k^- \mathbf{e}_k^{T-}] = E[(\mathbf{x}_k - \hat{\mathbf{x}}_k^-)(\mathbf{x}_k - \hat{\mathbf{x}}_k^-)^T] \quad (5.4.7)$$

In many cases, we begin the estimation problem with no prior measurements. Thus, in this case, if the process mean is zero, the initial estimate is zero, and the associated error covariance matrix is just the covariance matrix of \mathbf{x} itself.

With the assumption of a prior estimate $\hat{\mathbf{x}}_k^-$, we now seek to use the measurement \mathbf{z}_k to improve the prior estimate. We choose a linear blending of the noisy measurement and the prior estimate in accordance with the equation

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_k^-) \quad (5.4.8)$$

where

$\hat{\mathbf{x}}_k$ = updated estimate

\mathbf{K}_k = blending factor (yet to be determined)

The problem now is to find the particular blending factor \mathbf{K}_k that yields an updated estimate that is optimal in some sense. Just as in the Wiener solution, we use minimum mean-square error as the performance criterion. Toward this

*We tacitly assume here that the estimation error has zero mean, and thus it is proper to refer to $E[\mathbf{e}_k^- \mathbf{e}_k^{T-}]$ as a covariance matrix. It is also, of course, a moment matrix, but it is usually not referred to as such.

end we first form the expression for the error covariance matrix associated with the updated (a posteriori) estimate.

$$\mathbf{P}_k = E[\mathbf{e}_k \mathbf{e}_k^T] = E[(\mathbf{x}_k - \hat{\mathbf{x}}_k)(\mathbf{x}_k - \hat{\mathbf{x}}_k)^T] \quad (5.4.9)$$

Next, we substitute Eq. (5.4.2) into Eq. (5.4.8) and then substitute the resulting expression for $\hat{\mathbf{x}}_k$ into Eq. (5.4.9). The result is

$$\begin{aligned} \mathbf{P}_k = E\{ & [(\mathbf{x}_k - \hat{\mathbf{x}}_k^-) - \mathbf{K}_k(\mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k - \mathbf{H}_k \hat{\mathbf{x}}_k^-)] \\ & [(\mathbf{x}_k - \hat{\mathbf{x}}_k^-) - \mathbf{K}_k(\mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k - \mathbf{H}_k \hat{\mathbf{x}}_k^-)]^T \} \end{aligned} \quad (5.4.10)$$

Now, performing the indicated expectation and noting the $(\mathbf{x}_k - \hat{\mathbf{x}}_k^-)$ is the a priori estimation error that is uncorrelated with the measurement error \mathbf{v}_k , we have

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^- (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)^T + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^T \quad (5.4.11)$$

Notice here that Eq. (5.4.11) is a perfectly general expression for the updated error covariance matrix, and it applies for any gain \mathbf{K}_k , suboptimal or otherwise.

Returning to the optimization problem, we wish to find the particular \mathbf{K}_k that minimizes the individual terms along the major diagonal of \mathbf{P}_k , because these terms represent the estimation error variances for the elements of the state vector being estimated. The optimization can be done a number of ways. Our derivation here will follow the completing-the-square approach of Sorenson (6), because it does not involve the use of special matrix differentiation formulas which are required in the differential calculus approach (7). (Also, see Problem 5.10.) We now temporarily drop the subscripts in our equations in order to avoid unnecessary clutter in the derivation. Equation (5.4.11) is then rewritten without subscripts as

$$\mathbf{P} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^- (\mathbf{I} - \mathbf{K}\mathbf{H})^T + \mathbf{K}\mathbf{R}\mathbf{K}^T \quad (5.4.12)$$

This expression for \mathbf{P} may now be expanded and terms regrouped to yield

$$\mathbf{P} = \mathbf{P}^- - \underbrace{\mathbf{K}\mathbf{H}\mathbf{P}^-}_{\text{Linear in } \mathbf{K}} - \underbrace{\mathbf{P}^- \mathbf{H}^T \mathbf{K}^T}_{\text{Quadratic in } \mathbf{K}} + \underbrace{\mathbf{K}(\mathbf{H}\mathbf{P}^- \mathbf{H}^T + \mathbf{R})\mathbf{K}^T}_{\text{Quadratic in } \mathbf{K}} \quad (5.4.13)$$

\mathbf{P} can be seen to be quadratic in \mathbf{K} , and we now wish to do the matrix equivalent of “completing the square.” First we assume $(\mathbf{H}\mathbf{P}^- \mathbf{H}^T + \mathbf{R})$ to be symmetric and positive definite. It can then be written in factored form as $\mathbf{S}\mathbf{S}^T$, that is,

$$\mathbf{S}\mathbf{S}^T \triangleq \mathbf{H}\mathbf{P}^- \mathbf{H}^T + \mathbf{R} \quad (5.4.14)$$

(We will see presently that we do not need to find \mathbf{S} ; we only need to know this factored form is possible.) Using Eq. (5.4.14), the expression for \mathbf{P} may now be rewritten as

$$\mathbf{P} = \mathbf{P}^- - \mathbf{K}\mathbf{H}\mathbf{P}^- - \mathbf{P}^- \mathbf{H}^T \mathbf{K}^T + \mathbf{K}\mathbf{S}\mathbf{S}^T \mathbf{K}^T \quad (5.4.15)$$

Now we complete the square and write \mathbf{P} in the form

$$\mathbf{P} = \mathbf{P}^- + (\mathbf{K}\mathbf{S} - \mathbf{A})(\mathbf{K}\mathbf{S} - \mathbf{A})^T - \mathbf{A}\mathbf{A}^T \quad (5.4.16)$$

where \mathbf{A} does not involve \mathbf{K} . If Eq. (5.4.16) is expanded and compared term by term with Eq. (5.4.15), we see that the following equality must be true:

$$\mathbf{K}\mathbf{S}\mathbf{A}^T + \mathbf{A}\mathbf{S}^T\mathbf{K}^T = \mathbf{K}\mathbf{H}\mathbf{P}^- + \mathbf{P}^-\mathbf{H}^T\mathbf{K}^T \quad (5.4.17)$$

It is easily verified that if we let \mathbf{A} be

$$\mathbf{A} = \mathbf{P}^-\mathbf{H}^T(\mathbf{S}^T)^{-1} \quad (5.4.18)$$

then Eq. (5.4.17) is satisfied, and Eq. (5.4.16) is equivalent to Eq. (5.4.15).

We now note that the first and third terms in Eq. (5.4.16) do not involve \mathbf{K} . Only the middle term involves \mathbf{K} , and it is the product of a matrix and its transpose, which ensures that all terms along the major diagonal will be nonnegative. We wish to minimize the diagonal terms of \mathbf{P} . Clearly, then, the best we can possibly hope to do is to adjust \mathbf{K} to make the middle term of Eq. (5.4.16) zero. Thus, we choose \mathbf{K} to be such that

$$\mathbf{K}\mathbf{S} = \mathbf{A} \quad (5.4.19)$$

Or, using Eq. (5.4.18), we have

$$\begin{aligned} \mathbf{K} &= \mathbf{A}\mathbf{S}^{-1} \\ &= \mathbf{P}^-\mathbf{H}^T(\mathbf{S}^T)^{-1}\mathbf{S}^{-1} \\ &= \mathbf{P}^-\mathbf{H}^T(\mathbf{S}\mathbf{S}^T)^{-1} \end{aligned} \quad (5.4.20)$$

But $\mathbf{S}\mathbf{S}^T$ is just the factored form of $(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})$, so we have for our final expression for the optimum \mathbf{K} (with the subscripts reinserted):

$$\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)^{-1} \quad (5.4.21)$$

This particular \mathbf{K}_k , namely the one that minimizes the mean square estimation error, is called the *Kalman gain*.

The covariance matrix associated with the optimal estimate may now be computed. Referring to Eqs. (5.4.12) and (5.4.13), and reinserting the subscripts, we have

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \quad (5.4.22)$$

$$= \mathbf{P}_k^- - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^- - \mathbf{P}_k^-\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)\mathbf{K}_k^T \quad (5.4.23)$$

Routine substitution of the optimal gain expression, Eq. (5.4.21), into Eq. (5.4.23) leads to

$$\mathbf{P}_k = \mathbf{P}_k^- - \mathbf{P}_k^-\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)^{-1}\mathbf{H}_k\mathbf{P}_k^- \quad (5.4.24)$$

or

$$\mathbf{P}_k = \mathbf{P}_k^- - \mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)\mathbf{K}_k^T \quad (5.4.25)$$

or

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^- \quad (5.4.26)$$

Of the three expressions for \mathbf{P}_k , the latter one given by Eq. (5.4.26) is the simplest, so it is used more frequently than the others. Again, note that Eq. (5.4.22) is valid for any gain, suboptimal or otherwise, whereas Eqs. (5.4.24), (5.4.25), and (5.4.26) are valid only for the Kalman (optimal) gain.

We now have a means of assimilating the measurement at t_k by the use of Eq. (5.4.8) with \mathbf{K}_k set equal to the Kalman gain as given by Eq. (5.4.21). Note that we need $\hat{\mathbf{x}}_k^-$ and \mathbf{P}_k^- to accomplish this, and we can anticipate a similar need at the next step in order to make optimal use of the measurement \mathbf{z}_{k+1} . The updated estimated $\hat{\mathbf{x}}_k$ is easily projected ahead via the transition matrix. We are justified in ignoring the contribution of \mathbf{w}_k in Eq. (5.4.1) because it has zero mean and is uncorrelated with the previous \mathbf{w} 's. Thus we have

$$\hat{\mathbf{x}}_{k+1}^- = \boldsymbol{\phi}_k \hat{\mathbf{x}}_k \quad (5.4.27)$$

The error covariance matrix associated with $\hat{\mathbf{x}}_{k+1}^-$ is obtained by first forming the expression for the a priori error

$$\begin{aligned} \mathbf{e}_{k+1}^- &= \mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}^- \\ &= (\boldsymbol{\phi}_k \mathbf{x}_k + \mathbf{w}_k) - \boldsymbol{\phi}_k \hat{\mathbf{x}}_k \\ &= \boldsymbol{\phi}_k \mathbf{e}_k + \mathbf{w}_k \end{aligned} \quad (5.4.28)$$

We now note that \mathbf{w}_k and \mathbf{e}_k are uncorrelated, and thus we can write the expression for \mathbf{P}_{k+1}^- as

$$\begin{aligned} \mathbf{P}_{k+1}^- &= E[\mathbf{e}_{k+1}^-\mathbf{e}_{k+1}^{T-}] = E[(\boldsymbol{\phi}_k \mathbf{e}_k + \mathbf{w}_k)(\boldsymbol{\phi}_k \mathbf{e}_k + \mathbf{w}_k)^T] \\ &= \boldsymbol{\phi}_k \mathbf{P}_k \boldsymbol{\phi}_k^T + \mathbf{Q}_k \end{aligned} \quad (5.4.29)$$

We now have the needed quantities at time t_{k+1} , and the measurement \mathbf{z}_{k+1} can be assimilated just as in the previous step.

Equations (5.4.8), (5.4.21), (5.4.26), (5.4.27), and (5.4.29) comprise the Kalman filter recursive equations. It should be clear that once the loop is entered, it can be continued ad infinitum. The pertinent equations and the sequence of computational steps are shown pictorially in Fig. 5.7. This summarizes what is now known as the *Kalman filter*.

Before proceeding to some examples, it is interesting to reflect on the Kalman filter in perspective. If you were to stumble onto the recursive process of Fig. 5.7 without benefit of previous history, you might logically ask, "Why in the world did somebody call that a filter? It looks more like a computer algorithm." You would, of course, be quite right in your observation. The Kalman filter is just a computer algorithm for processing discrete measurement data in an optimal fashion. Its roots, though, go back to the days when filters were made of electrical elements wired together in such a way as to yield the

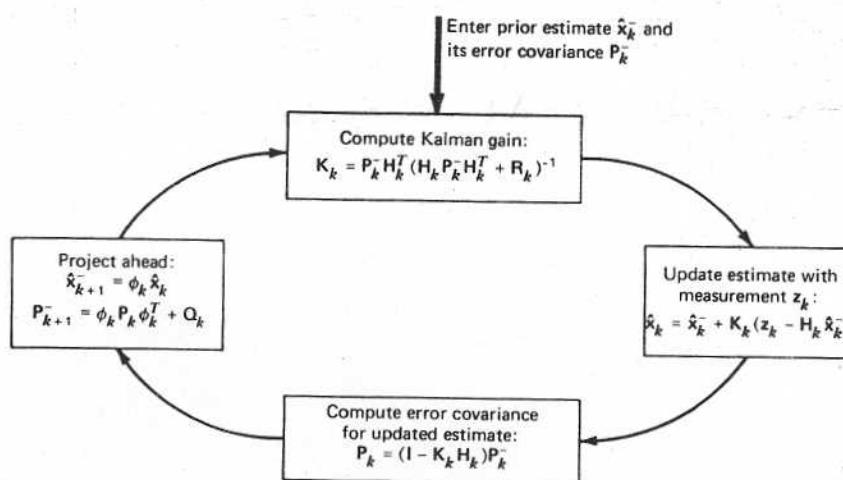


Figure 5.7. Kalman filter loop.

desired frequency response. The design was often heuristic. Wiener then came on the scene in the 1940s and added a more sophisticated type of filter problem. The end result of his solution was a filter weighting function or a corresponding transfer function in the complex domain. Implementation in terms of electrical elements was left as a further exercise for the designer. The sampled-data version of the Wiener problem remained unsolved (in a practical sense, at least) until Kalman's paper of 1960. Even though his presentation appeared to be quite abstract at first glance, engineers soon realized that his work provided a practical solution to a number of unsolved filtering problems, especially in the field of navigation. More than 20 years have elapsed since Kalman's original paper, and there are still numerous current papers dealing with new applications and variations on the basic Kalman filter. It has withstood the test of time!

5.5 EXAMPLES

The basic recursive equations for the Kalman filter were presented in Section 5.4. Two simple scalar examples illustrate the use of these equations. Examples of higher-order systems will be deferred until Chapter 6.

Example 1. Wiener (Brownian motion) Process Consider the Wiener process described in Fig. 5.8. Recall that this process has Gaussian statistics and is assumed to be zero at $t = 0$. Assume that we have a sequence of independent noisy measurements taken at unit intervals as shown in Fig. 5.8 and let the standard deviation of the measurement error be one-half. We wish

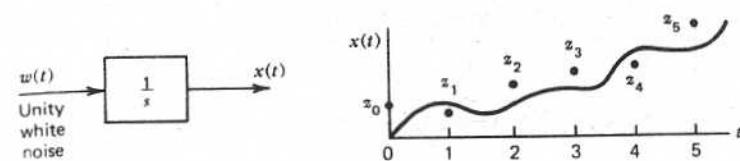


Figure 5.8. Block diagram of Wiener process and typical sample function of the process.

to determine the optimal estimate of the process at the sample times 0, 1, 2, 3, ..., etc.

The model parameters for the Kalman filter may be computed as follows: The process differential equation for this case is

$$\dot{x} = w(t) \quad (5.5.1)$$

Therefore, the transition matrix is

$$\phi_k = 1 \quad (5.5.2)$$

The Q_k matrix is computed as

$$\begin{aligned} Q_k &= E[w_k w_k] = E\left[\int_0^1 1 \cdot w(u) du \int_0^1 1 \cdot w(v) dv\right] \\ &= \int_0^1 \int_0^1 E[w(u)w(v)] du dv = \int_0^1 \int_0^1 \delta(u-v) du dv = 1 \end{aligned} \quad (5.5.3)$$

Since the measurement has a direct one-to-one correspondence to the process $x(t)$, the H_k matrix is

$$H_k = 1 \quad (5.5.4)$$

The variance of the measurement error is

$$R_k = (\text{Standard deviation})^2 = (\frac{1}{2})^2 = \frac{1}{4} \quad (5.5.5)$$

Our initial estimate at $t = 0$ is \hat{x}_0^- and is zero because the Wiener process is defined to begin at zero. Furthermore, because of this prior knowledge about the process, the error associated with our initial estimate is zero; that is, the a priori estimate, $\hat{x}_0^- = 0$, is perfect by definition of the process. Thus

$$P_0^- = 0 \quad (5.5.6)$$

We now have all the parameters needed for the Kalman filter and are ready to begin the recursive process. Referring to Fig. 5.7, we enter the loop at $t = 0$ and process the first measurement.

Step 1: $t = 0$ (Subscripts will be dropped for constant parameters.)

$$\begin{aligned} \text{Compute gain: } K_0 &= P_0^- H^T (H P_0^- H^T + R)^{-1} \\ &= 0/0.25 = 0 \end{aligned} \quad (5.5.7)$$

$$\begin{aligned} \text{Update estimate: } & \hat{x}_0 = \hat{x}_0^- + K_0(z_0 - H\hat{x}_0^-) \\ & = \hat{x}_0^- = 0 \end{aligned} \quad (5.5.8)$$

$$\begin{aligned} \text{Update } P: & P_0 = (I - K_0 H)P_0^- \\ & = 1 \cdot P_0^- = 0 \end{aligned} \quad (5.5.9)$$

$$\begin{aligned} \text{Project ahead: } & \hat{x}_1^- = \phi\hat{x}_0 = 1 \cdot 0 = 0 \\ & P_1^- = \phi P_0 \phi^T + Q \\ & = 1 \cdot 0 \cdot 1 + 1 = 1 \end{aligned} \quad (5.5.10) \quad (5.5.11)$$

Note that the measurement at $t = 0$ is given zero weight relative to the prior estimate. This makes good sense because the initial estimate is known to be perfect, whereas the measurement is known to be noisy.

Step 2: $t = 1$

$$\begin{aligned} \text{Compute gain: } & K_1 = P_1^- H^T (H P_1^- H^T + R)^{-1} \\ & = 1 \cdot 1 (1 \cdot 1 \cdot 1 + \frac{1}{4})^{-1} = \frac{4}{5} \end{aligned} \quad (5.5.12)$$

$$\begin{aligned} \text{Update estimate: } & \hat{x}_1 = \hat{x}_1^- + K_1(z_1 - H\hat{x}_1^-) \\ & = 0 + \frac{4}{5}(z_1) \end{aligned} \quad (5.5.13)$$

$$\begin{aligned} \text{Update } P: & P_1 = (I - K_1 H)P_1^- \\ & = (1 - \frac{4}{5} \cdot 1)1 = \frac{1}{5} \end{aligned} \quad (5.5.14)$$

$$\begin{aligned} \text{Project ahead: } & \hat{x}_2^- = \phi\hat{x}_1 = 1 \cdot \hat{x}_1 \\ & P_2^- = \phi P_1 \phi^T + Q \\ & = 1 \cdot \frac{1}{5} \cdot 1 + 1 = \frac{6}{5} \end{aligned} \quad (5.5.15) \quad (5.5.16)$$

This could now be carried on ad infinitum. Note that with step 2 the filter begins to give the measurement some weight in determining the optimal estimate. As times goes on, the filter depends more and more on the measurements and less on the initial assumptions. ■

Example 2. Scalar Gauss-Markov Process Consider a stationary Gauss-Markov process whose autocorrelation function is

$$R_x(\tau) = 1 \cdot e^{-|\tau|} \quad (5.5.17)$$

Clearly, the correlation time and variance for this process are both unity. Therefore, the spectral function for the process is

$$S_x(s) = \frac{2}{-s^2 + 1} = \frac{\sqrt{2}}{s+1} \cdot \frac{\sqrt{2}}{-s+1} \quad (5.5.18)$$

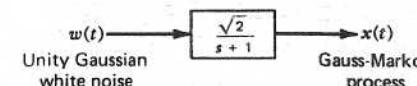


Figure 5.9. Shaping filter for Gauss-Markov Process.

and the shaping filter that shapes white noise into the process is shown in Fig. 5.9. Thus, the state equation for this process is

$$\dot{x} = -x + \sqrt{2}w(t) \quad (5.5.19)$$

Suppose we have a sequence of noisy measurements of this process taken .02 sec apart beginning at $t = 0$. The measurement error will be assumed to have a variance of unity. We wish to process these via a Kalman filter and obtain an optimal estimate of $x(t)$. First we need to determine the filter parameters ϕ_k , H_k , Q_k , and R_k .

The state transition matrix (scalar in this case):

$$\phi_k = e^{-.02} \approx .9802 \quad (5.5.20)$$

The measurement relationship to x :

$$H_k = 1 \quad (5.5.21)$$

The input noise sequence:

$$\begin{aligned} Q_k &= E[w_k^2] = E\left[\int_0^{.02} \sqrt{2} e^{-u} w(u) du \int_0^{.02} \sqrt{2} e^{-v} w(v) dv\right] \\ &= \int_0^{.02} (\sqrt{2} e^{-v})^2 dv = 1 - e^{-2(.02)} \approx .03921 \end{aligned} \quad (5.5.22)$$

The measurement error:

$$R_k = 1 \quad (5.5.23)$$

We also need the initial conditions \hat{x}_0^- and P_0^- to enter the recursive loop. In this case, we have assumed that the process is Gauss-Markov and stationary with a known autocorrelation function. We have no measurements prior to $t = 0$, but the assumed knowledge of the process autocorrelation function tells us the process has zero mean and a variance of unity. This is important information and enables us to start the recursive process at $t = 0$ with initial conditions

$$\hat{x}_0^- = 0 \quad (5.5.24)$$

$$P_0^- = 1 \quad (5.5.25)$$

Now that the four filter parameters and initial conditions are determined, it is a routine matter to cycle through the Kalman filter loop shown in Fig. 5.7 as many times as desired.

In order to add a note of realism to this example, the Markov process and noisy measurement situation just described were simulated using random num-

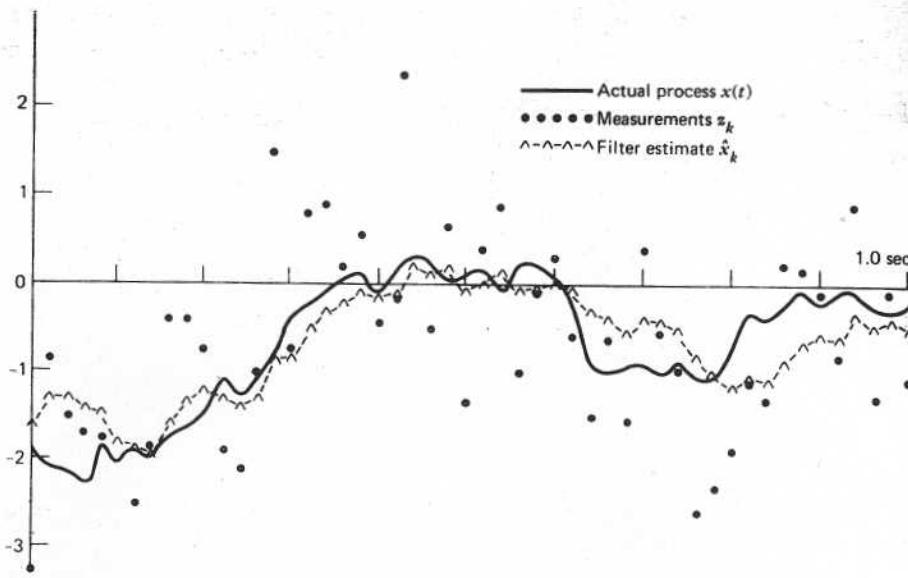


Figure 5.10. Simulation for Gauss-Markov example.

bers selected by chance. The results for the first 50 steps of the simulation are shown in Fig. 5.10. The discrete measurements z_k are shown as dots, and it should be obvious that we have postulated a very noisy measurement situation for this example. In spite of this, though, the filter does a reasonably good job of tracking $x(t)$. After about the first 20 steps, the filter settles down to a steady-state condition where the Kalman gain is about .165 and the standard deviation of the filter error is about .4. That is, $\sqrt{P_k}$ in the steady state is about .4 units. This is consistent, qualitatively at least, with what we see in the results shown in Fig. 5.10. It is comforting to know that our single sample realization of the process is typical rather than atypical. There is always a risk, of course, that a single trial will produce the unusual event rather than the usual one. ■

5.6 THE GAUSSIAN PROCESS AND CONDITIONAL DENSITY VIEWPOINT

In our discussion of optimization thus far, we have used minimum mean-square error as the performance criterion. This was done partly as a matter of mathematical convenience, but not entirely so because minimizing the mean-square error forces the filter to be optimal by other standards as well, especially in the Gaussian case. We now elaborate on this.

We begin by assuming the process to be estimated and the measurement relation are in the usual form of Eqs. (5.4.1) and (5.4.2) and that we have Gaussian statistics throughout. We wish now to consider two random variables and their associated probability density functions. Let

x^* = the process random variable at time t_k , conditioned on all the prior measurement information up through t_{k-1} .

z = a random variable related to x^* via the equation

$$z = Hx^* + v_k \quad (5.6.1)$$

where v_k is $N(0, R)$ and independent of x^* .

Since all the probability densities being considered are normal, it is convenient to use the notation $N(\mathbf{m}, \mathbf{C})$ to indicate normal with mean \mathbf{m} and covariance \mathbf{C} . We now have the following probability densities for x^* and z :

$$f_{x^*} \sim N(\mathbf{m}^*, \mathbf{P}^*) \quad (5.6.2)$$

$$f_z \sim N(\mathbf{m}_z, \mathbf{C}_z) \quad (5.6.3)$$

where we assume that \mathbf{m}^* and \mathbf{P}^* are known from previous steps, and \mathbf{m}_z and \mathbf{C}_z are given by

$$\mathbf{m}_z = H\mathbf{m}^* \quad (5.6.4)$$

$$\mathbf{C}_z = H\mathbf{P}^*H^T + R \quad (5.6.5)$$

Equation (5.6.4) comes from Eq. (5.6.1) and the assumption that the mean of v_k is zero. Equation (5.6.5) also comes from the linear connection between x^* and z , and it is derived by simply forming the expectation of zz^T .

Equation (5.6.1) also enables us to write one of the needed conditional densities simply by inspection:

$$f_{z|x^*} \sim N(Hx^*, R) \quad (5.6.6)$$

The other conditional density function may now be found from Bayes relation

$$f_{x^*|z} = \frac{f_{z|x^*} f_{x^*}}{f_z} \quad (5.6.7)$$

Recall now that x^* is the random variable x_k conditioned on all measurements up through z_{k-1} . Also, up to this point z is simply another random variable defined by Eq. (5.6.1). Now think of z as assuming a particular realization z_k , that is, the measurement at t_k . Then $f_{x^*|z}$ becomes the density of x_k conditioned on all measurements up through z_k .

In visualizing the various probability densities here, it may help to think of a conceptual statistical experiment. Imagine a large number of chance trials of the x process and associated measurements up through time t_k . If we simply look at the statistical distribution of all the resulting x_k 's we find they are

unbiased with some covariance \mathbf{C}_x . Now, out of the mass of trials, sort out just those results corresponding to particular realizations $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{k-1}$. This will yield a subset of \mathbf{x}_k 's that are approximately distributed as $N(\mathbf{m}^*, \mathbf{P}^*)$. Now, if we further sort the subset and sort out just those trials yielding \mathbf{z}_k , we have a set of \mathbf{x}_k 's with an even different distribution—that given by $f_{x^*|z_k}$ [Eq. (5.6.7)] where $\mathbf{z} = \mathbf{z}_k$. We now wish to look at the mean and covariance of this distribution.

Each density function on the right side of Eq. (5.6.7) is normal and may be written out explicitly in the standard exponential form (see Section 1.15). The result may then also be algebraically reduced to the standard form. This involves considerable algebraic effort, and so it will be left as an exercise (see Problem 5.11). The final result is

$$f_{x^*|z} \sim N(\mathbf{m}, \mathbf{C}) \quad (5.6.8)$$

where

$$\mathbf{m} = \mathbf{m}^* + \mathbf{P}^* \mathbf{H}^T (\mathbf{H} \mathbf{P}^* \mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{z} - \mathbf{H} \mathbf{m}^*) \quad (5.6.9)$$

and

$$\mathbf{C} = [(\mathbf{P}^*)^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^{-1} \quad (5.6.10)$$

The expression for the mean \mathbf{m} has the following significance relative to the estimation problem. If we choose as our best estimate the process mean conditioned on all variable measurement data, then the expression given by Eq. (5.6.9) is identical to that obtained by minimizing the mean-square error. We simply replace \mathbf{m}^* and \mathbf{P}^* in Eq. (5.6.9) with $\hat{\mathbf{x}}^-$ and \mathbf{P}^- and we have Eq. (5.4.8) with \mathbf{K}_k set equal to the Kalman gain. The expression for the covariance given by Eq. (5.6.10) may not look familiar, but in Chapter 6 it will be shown to be identically equal to the $\mathbf{P} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^-$ expression, provided \mathbf{K} is the Kalman gain.

There are some far-reaching consequences of the conditional mean viewpoint:

1. For the Gaussian case, the conditional mean is also the "most likely" value in that the maximum of the density function occurs at the mean. Also, it can be shown that the conditional mean minimizes the expectation of most any reasonable nondecreasing function of the magnitude of the error (as well as the squared error). [See Meditch (8) for a more complete discussion of this.] Thus, in the Gaussian case, the Kalman filter is best by most any reasonable criterion.
2. We note that in the conditional density function approach, we did not need to assume a linear relationship between the estimate and the measurements. Instead, this came out naturally as a consequence of the Gaussian assumption and our choice of the conditional mean as our estimate. Thus, in the

Gaussian case, we know that we need not search among nonlinear filters for a better one; it cannot exist. Thus, our linear assumption in the derivation of both the Wiener and Kalman filters turns out to be a fortuitous one. That is, in the Gaussian case, the Wiener/Kalman filter is not just best within a class of linear filters; it is best within a class of all filters, linear or nonlinear.

3. In physical problems, we often begin with incomplete knowledge of the process under consideration. Perhaps only the covariance structure of the process is known. In this case we can always imagine a corresponding Gaussian process with the same covariance structure. This process is then completely defined and conclusions for the equivalent Gaussian process can be drawn. It is, of course, a bit risky to extend these conclusions to the original process, not knowing it to be Gaussian. However, even risky conclusions are better than none if viewed with proper caution.

Problems

- 5.1. A stationary Gaussian random process is known to have a spectral density function of the form

$$S_y(j\omega) = \frac{\omega^2 + 1}{\omega^4 + 8\omega^2 + 16}$$

Assume that discrete noisy measurements are available at $t = 0, 1, 2, 3, \dots$ and that the measurement errors are uncorrelated and have a variance of two units. Model this process in appropriate form for a Kalman filter. That is, find ϕ_k , \mathbf{H}_k , \mathbf{Q}_k , \mathbf{R}_k and the initial conditions $\hat{\mathbf{x}}_0^-$ and \mathbf{P}_0^- . Assume there is no measurement information prior to $t = 0$.

- 5.2. A classical problem in Wiener-filter theory is one of separating signal from noise when both the signal and noise have exponential autocorrelation functions. Let the noisy measurement be

$$z(t) = s(t) + n(t)$$

and let signal and noise be stationary independent processes with autocorrelation functions

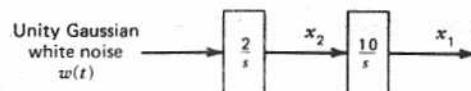
$$R_s(\tau) = \sigma_s^2 e^{-\beta_s |\tau|}$$

$$R_n(\tau) = \sigma_n^2 e^{-\beta_n |\tau|}$$

Assume we have discrete samples of $z(t)$ spaced Δt apart and wish to form the optimal estimate of $s(t)$ at the sample points. Let $s(t)$ and $n(t)$ be the first and second elements of the process state vector, and then find the parameters of the Kalman filter for this situation. That is, find ϕ_k , \mathbf{H}_k , \mathbf{Q}_k , \mathbf{R}_k and the initial conditions $\hat{\mathbf{x}}_0^-$ and \mathbf{P}_0^- . Assume the measurement sequence begins at $t = 0$.

- 5.3. Consider the two cascaded integrators as shown in the accompanying figure. Integrator scale factors are shown in the "boxes." The initial conditions for both integrators are zero, and state variables x_1 and x_2 are defined as shown on the figure. State variable x_1 is the quantity being measured and it is sampled at 2-second intervals. The

measurement error variance is four units. Find the Kalman filter parameters ϕ_k , \mathbf{H}_k , \mathbf{Q}_k , and \mathbf{R}_k and the appropriate initial conditions $\hat{\mathbf{x}}_0^-$, \mathbf{P}_0^- for this situation.



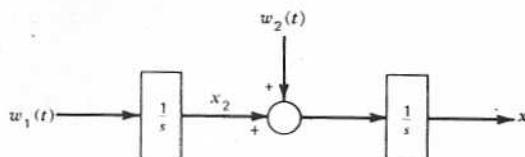
Problem 5.3

5.4. The system shown is driven by two independent Gaussian white sources $w_1(t)$ and $w_2(t)$. Their spectral functions are given by

$$S_{w_1} = 4 \text{ (ft/sec}^2\text{)}^2/\text{rad/sec}$$

$$S_{w_2} = 16 \text{ (ft/sec}^2\text{)}^2/\text{rad/sec}$$

Let state variables be chosen as shown on the diagram, and assume noisy measurements of x_1 are obtained at unit intervals of time. A discrete Kalman filter model is desired. Find \mathbf{Q}_k for this model.



Problem 5.4

5.5. In Example 1, Section 5.5, the random walk process was assumed to have an initial value of zero, that is, it was a Wiener process. Consider a similar process except that the initial value is a random variable described as $N(0, 1)$.

- (a) How will this change in initial condition affect the discrete Kalman filter model?
- (b) Cycle through two recursive steps and find the Kalman gain at $t = 1$. Compare this with the corresponding gain in Example 1.
- (c) Would you expect the two gains, that is, for Example 1 and this problem, to approach the same value as the number of steps becomes large? (If in doubt, cycle through a few more steps for both cases.)

5.6. Consider the same Gauss-Markov process and measurement situation discussed in the second example of Section 5.5.

- (a) Carry out two recursive steps of the Kalman filter and write the estimate at $t = .02$ sec explicitly in terms of z_0 and z_1 .
- (b) Using the weighting function approach of Section 4.8, write the optimal estimate after two measurements (i.e., at $t = .02$ sec) in terms of z_0 and z_1 . [Note that the answer should be the same as that obtained in part (a).]

5.7. A common process encountered in physical applications is harmonic motion with random amplitude and phase. If the motion is truly harmonic, it satisfies the undriven differential equation

$$\ddot{x} + \omega_r^2 x = 0$$

and it can be seen to be a deterministic random process. Suppose we have a sequence of discrete noisy measurements of x beginning at $t = 0$, and we wish to estimate x . Assume

that we have no prior knowledge of either amplitude or phase, and that the measurements are equally spaced by Δt and each has a measurement variance of R .

- (a) Choose phase variables (i.e., x and \dot{x}) as the state variables and find the Kalman filter parameters for this situation. (*Partial Answer:* In the absence of better information, it is appropriate to let $\hat{\mathbf{x}}_0^- = \mathbf{0}$ and $\mathbf{P}_0^- = \infty$. As a practical matter, \mathbf{P}_0^- can be set at a very large value relative to the measurement variance and thus the same result as “ ∞ ” is accomplished, approximately. There is also an alternative approach that is discussed in Chapter 6.)
- (b) It will be noted that the solution for x is of the form

$$x(t) = x(0) \cos \omega_r t + \frac{\dot{x}(0)}{\omega_r} \sin \omega_r t$$

This suggests that we might choose $x(0)$ and $\dot{x}(0)$ as our state variables and then model the Kalman filter accordingly. Work out the filter parameters for this model.

- (c) Find the linear transformation relating the state variables of parts (a) and (b). (*Note:* The elements of the transformation matrix will be time-varying in this case.)

5.8. Suppose that we make the following linear transformation on the process state vector of Problem 5.2:

$$\begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

This transformation is nonsingular, and hence we should be able to consider \mathbf{x}' as the state vector to be estimated and write out the Kalman filter equations accordingly. Specify the Kalman filter parameters for the transformed problem. (Note that the specified transformation yields a simplification of the measurement matrix, but this is at the expense of complicating the model elsewhere.)

5.9. It is almost self-evident that if the estimation errors are minimized in one set of state variables, this also will minimize the error in any linear combination of those state variables. This can be shown formally by considering a new state vector \mathbf{x}' to be related to the original state vector \mathbf{x} via a general nonsingular transformation $\mathbf{x}' = \mathbf{A}\mathbf{x}$. Proceeding in this manner, show that the Kalman estimate obtained in the transformed domain is the same as would be obtained by performing the update [i.e., Eq. (5.4.8)] in the original \mathbf{x} domain and then transforming this estimate via the \mathbf{A} matrix.

5.10. The equation for Kalman gain, Eq. (5.4.21), may also be derived using differential calculus. In order to do this, the general expression for \mathbf{P} , Eq. (5.4.13), will be repeated here for convenience.

$$\mathbf{P} = \mathbf{P}^- - \mathbf{K}\mathbf{H}\mathbf{P}^- - \mathbf{P}^-\mathbf{H}^T\mathbf{K}^T + \mathbf{K}(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})\mathbf{K}^T$$

The elements along the major diagonal of \mathbf{P} are to be minimized with respect to \mathbf{K} . We first note that this can be accomplished by minimizing the sum of the elements which is known as the trace of \mathbf{P} , that is, $\text{trace } \mathbf{P} = p_{11} + p_{22} + \dots + p_{nn}$. Also, two general

matrix differentiation formulas are helpful in the minimization:

$$(1) \frac{d[\text{trace}(\mathbf{AB})]}{d\mathbf{A}} = \mathbf{B}^T, \quad (\mathbf{AB} \text{ must be square})$$

$$(2) \frac{d[\text{trace}(\mathbf{ACA}^T)]}{d\mathbf{A}} = 2\mathbf{AC}, \quad (\mathbf{C} \text{ must be symmetric})$$

where the derivative of a scalar s with respect to a matrix \mathbf{A} is defined as

$$\frac{ds}{d\mathbf{A}} = \begin{bmatrix} \frac{ds}{da_{11}} & \frac{ds}{da_{12}} & \dots \\ \frac{ds}{da_{21}} & \frac{ds}{da_{22}} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Formulas (1) and (2) can be verified by writing out a few terms of the indicated traces and then performing the specified scalar differentiations. Beginning with the given expression for \mathbf{P} , find the optimal \mathbf{K} using differential calculus methods.

5.11. Show that the mean and covariance associated with conditional density function $f_{x^*|z}$ are as given by Eqs. (5.6.9) and (5.6.10) (*Hint:* When the expression for $f_{x^*|z}$ is written out explicitly in terms of x^* , it will be found that the quantity in the exponent is quadratic in x^* . Next, you will need to do the matrix equivalent of completing-the-square in order to recognize the mean and covariance of the density function. Also, you will find the alternative gain and error-covariance expressions derived in Chapter 6 to be useful here. They are, with subscripts omitted,

$$\mathbf{K} = \mathbf{PH}^T \mathbf{R}^{-1}$$

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

Assume these to be valid and use them wherever they may be helpful in this problem.)

5.12. It is desired to simulate a Wiener process at discrete sample points separated by .1 sec. Call the sequence of samples of the Wiener process x_0, x_1, x_2, \dots with x_0 being the sample at $t = 0$. The Wiener process under consideration may be thought of as integrated unity white noise. To accomplish the simulation, one has available a sequence of samples of a normal distribution $N(0, 1)$. Call this sequence u_0, u_1, \dots and write the difference equation for generating the x_0, x_1, x_2, \dots sequence from the u_0, u_1, \dots sequence.

5.13. The Gauss-Markov example of Section 5.5 is to be simulated again for 50 steps using the Gaussian random numbers accompanying this problem. Use the first 50 numbers (appropriately scaled) for the w_k sequence. Use the next 51 for the v_k sequence, and use the last number for the initial value of the x_k process. Since the first simulated measurement occurs at $t = 0$ and the last at $t = 1$, there are 51 simulated measurements in all. The algebra is scalar in this example, and so the recursive steps are easily programmed; or, with modest effort the calculations can be done manually with a handheld calculator. Plot the results in the same form as in Fig. 5.10, and state whether you think the results of this simulation are typical or atypical.

102 Samples from a Normal Distribution $N(0, 1)$

-1.706	.518	.737	.048	-.716	-1.728
.410	-.153	-.128	.649	-.927	-.493
.703	.197	.075	-1.477	.916	
-.923	.131	.848	-.182	-1.386	
.890	-.837	-.402	.668	-.919	
-.202	2.511	.365	.210	-.559	
.995	-.114	1.143	.628	-1.137	
.810	-.005	1.115	.285	.987	
-1.598	-.355	.703	1.055	.332	
-.235	-.967	1.080	-.105	1.190	
.018	-.233	.968	.650	-.920	
-.215	1.675	1.392	-.068	.348	
-.121	-.842	-.610	.683	.510	
-.128	.977	1.675	-.027	.719	
.974	-.365	-.817	-1.360	-.623	
1.911	-.668	-1.685	-.713	-.017	
.367	-1.054	-.451	-1.599	-1.160	
-.194	-.997	-.572	.752	1.927	
1.175	-.403	.144	.075	1.041	
-.593	1.190	-.028	-.156	-.105	

References Cited

1. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Trans. of the ASME—J. of Basic Engr.*, 35–45 (March 1960).
2. J. J. D'Azzo and C. H. Houpis, *Linear Control System Analysis and Design*, Second Edition, New York: McGraw-Hill, 1981.
3. R. C. Dorf, *Modern Control Systems*, 3rd ed., Reading, Mass.: Addison-Wesley, 1980.
4. T. A. Stansell, Jr., "The Many Faces of Transit," *Navigation*, Jour. of the Inst. of Navigation, 25: 1, 55–70 (Spring 1978).
5. P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, New York: Wiley, 1962.
6. H. W. Sorenson, Kalman Filtering Techniques, in *Advances in Control Systems* (Vol. 3), C. T. Leondes (Ed.), New York: Academic Press, 1966, pp. 219–289.
7. A. Gelb (Ed.), *Applied Optimal Estimation*, Cambridge, Mass.: MIT Press, 1974.
8. J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*, New York: McGraw-Hill, 1969.

Additional References on Kalman Filtering

9. P. S. Maybeck, *Stochastic Models, Estimation and Control* (Vol. 1), New York: Academic Press, 1979.
10. A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communications and Control*, New York: McGraw-Hill, 1971.

11. S. M. Bozic, *Digital and Kalman Filtering*, London: E. Arnold, Publisher, 1979.
12. B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Englewood Cliffs, N.J.: Prentice-Hall, 1979.
13. C. T. Leondes, (Ed.), *Theory and Application of Kalman Filtering*, North Atlantic Treaty Organization AGARD Report No. 139 (February 1970).
14. H. W. Sorenson, *Parameter Estimation*, New York: Marcel Dekker, 1980.

CHAPTER 6

Applications and Additional Topics on Discrete Kalman Filtering

In the period immediately following Kalman's original work, many extensions and variations were developed which enhanced the usefulness of the technique in applied work. This chapter continues the subject of discrete Kalman filtering with some of the more important extensions and related topics. Also a limited number of applications are presented to illustrate the versatility of Kalman filtering. The treatment of applications here must be brief. However, there is no shortage of papers on applications in a wide variety of fields. (For a start, see References 5, 11, and 20. These three volumes alone contain a wealth of examples.)

6.1 AUGMENTING THE STATE VECTOR

The Kalman filter is especially sensitive to format in that the process to be estimated and the measurement relationship must be of the exact form

$$\mathbf{x}_{k+1} = \boldsymbol{\phi}_k \mathbf{x}_k + \mathbf{w}_k \quad (6.1.1)$$

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \quad (6.1.2)$$

where \mathbf{w}_k and \mathbf{v}_k must be white (uncorrelated) sequences. We are assured that \mathbf{w}_k will be white if Eq. (6.1.1) is the sampled version of a continuous linear system driven by white noise. However, more often than not, the driving functions in the original problem are not white, and they, in turn, must be modeled with differential equations relating them to fictitious white noise processes. This creates additional state variables that must be appended to the original ones in the state model. An example will illustrate this.

Example 6.1 This example is due to B. E. Bona and R. J. Smay (1), and it was one of the early on-line applications of Kalman filtering. In marine applications, the mission time is usually long, and the ship's inertial navigation system is expected to operate continuously for weeks with a minimum of attention. One of the principle sources of errors in inertial systems is gyro drift, and this causes the position error to grow with time. Thus, the system must be reset occasionally with position fixes from other sources such as LORAN or a satellite navigation system, or any of a multitude of navigation aids. In addition to

resetting the position error, it is also desirable to correct the underlying sources giving rise to the error—the so-called gyro biases. These cannot be observed directly at sea, so one must be content only to observe them indirectly by observing their effect on the system position error.

The applicable error propagation equations for a damped inertial navigation system in a slow-moving vehicle are*

$$\dot{\psi}_x - \Omega_z \psi_y = \varepsilon_x \quad (6.1.3)$$

$$\dot{\psi}_y + \Omega_z \psi_x - \Omega_x \psi_z = \varepsilon_y \quad (6.1.4)$$

$$\dot{\psi}_z + \Omega_x \psi_y = \varepsilon_z \quad (6.1.5)$$

where x , y , and z denote the platform (gyro and accelerometer instrument cluster) coordinate axes in the north, west, and up directions, and

ψ_x = inertial system's west position error (in terms of angular measure in radians)

ψ_y = inertial system's south position error (in terms of angular measure in radians)

ψ_z = [platform azimuth error] − [west position error] · [tan (latitude)]

Also

Ω_x = x component of earth rate Ω [i.e., $\Omega_x = \Omega \cos (\text{Lat.})$]

Ω_z = z component of earth rate Ω [i.e., $\Omega_z = \Omega \sin (\text{Lat.})$]

and

ε_x , ε_y , ε_z = gyro drift rates for the x , y , and z axis gyros.

We assume that the ship's latitude is known approximately; therefore Ω_x and Ω_z are known constants.

The three differential equations [(6.1.3) to (6.1.5)] represent a third-order linear system with ε_x , ε_y , and ε_z as the driving functions. It is not reasonable to assume these to be white noises in this situation. Quite to the contrary, instead of "jumping around wildly" like white noise, just the reverse is to be expected. Gyro drift is due to mechanical causes that produce minute random torques about the gyro's output axis, and these usually change slowly with time. In the referenced paper, Bona and Smay suggest that the respective gyro drift rates may be properly modeled as a sum of a random constant plus a Markov component with a large time constant. This model accommodates slowly varying

* Equations (6.1.3) to (6.1.5) are certainly not obvious, and a considerable amount of background in inertial navigation theory is needed to understand the assumptions and approximations leading to this simple set of equations (2.3). We do not attempt to derive the equations here. For purposes of understanding the Kalman filter, simply assume that these equations do, in fact, accurately describe the error propagation in this application and proceed on to the details of the Kalman filter. (The natural modes of oscillation for this system of equations are discussed briefly in Problem 6.1.)

changes in the "bias" as well as a constant offset. Thus, we let

$$\begin{aligned} \varepsilon_x &= \varepsilon_{xm} + \varepsilon_{xc} \\ \varepsilon_y &= \varepsilon_{ym} + \varepsilon_{yc} \\ \varepsilon_z &= \varepsilon_{zm} + \varepsilon_{zc} \end{aligned} \quad (6.1.6)$$

where subscripts m and c denote Markov and constant components, respectively. Now, from Section 5.2 we know that random constants and Markov processes can be modeled with simple first-order differential equations. Thus, we have for the Markov components

$$\begin{aligned} \dot{\varepsilon}_{xm} &= -\beta_x \varepsilon_{xm} + w_x \\ \dot{\varepsilon}_{ym} &= -\beta_y \varepsilon_{ym} + w_y \\ \dot{\varepsilon}_{zm} &= -\beta_z \varepsilon_{zm} + w_z \end{aligned} \quad (6.1.7)$$

where w_x , w_y , and w_z are independent white noise processes. For the constant components, we have

$$\begin{aligned} \dot{\varepsilon}_{xc} &= 0 \\ \dot{\varepsilon}_{yc} &= 0 \\ \dot{\varepsilon}_{zc} &= 0 \end{aligned} \quad (6.1.8)$$

We now have six additional first-order differential equations that must be appended to the original three equations. The addition of the new state variables is often called "augmenting the state vector." Regardless of terminology, the system is now ninth order. Let

$$\begin{aligned} \psi_x, \psi_y, \psi_z &= \text{state variables } x_1, x_2, x_3 \\ \varepsilon_{xm}, \varepsilon_{ym}, \varepsilon_{zm} &= \text{state variables } x_4, x_5, x_6 \\ \varepsilon_{xc}, \varepsilon_{yc}, \varepsilon_{zc} &= \text{state variables } x_7, x_8, x_9 \end{aligned}$$

The complete state model is then

$$\begin{array}{c|ccc|c|cc|c|c} \dot{x}_1 & 0 & \Omega_z & 0 & \mathbf{I} & & \mathbf{I} & x_1 & 0 \\ \dot{x}_2 & -\Omega_z & 0 & \Omega_x & & & & x_2 & 0 \\ \dot{x}_3 & 0 & -\Omega_x & 0 & & & & x_3 & 0 \\ \hline \dot{x}_4 & & & & -\beta_x & 0 & 0 & x_4 & w_x \\ \dot{x}_5 & & & & 0 & -\beta_y & 0 & x_5 & w_y \\ \dot{x}_6 & & & & 0 & 0 & -\beta_z & x_6 & w_z \\ \hline \dot{x}_7 & & & & & & & x_7 & 0 \\ \dot{x}_8 & & & & & & & x_8 & 0 \\ \dot{x}_9 & & & & & & & x_9 & 0 \end{array} \quad (6.1.9)$$

Note that the forcing functions appearing in the continuous model are now white; hence we are assured that the discrete model, that is, the solution of Eq. (6.1.9), will be of the proper form given by Eq. (6.1.1). Once the parameters in the continuous model are set, it is a routine matter to determine ϕ_k and Q_k for any time interval (t_k, t_{k+1}) using the methods given in Section 5.3. It is assumed that the continuous-model parameters are known from physical considerations of the particular equipment at hand.

The measurements in this application will be a sequence of position fixes obtained from sources completely independent of the inertial system that is being reset. The difference between the independently determined position and the inertial system's output will yield noisy measurements of the inertial *N-S* and *E-W* position errors; the noise in this case being the error associated with external reference providing the discrete position fix. Presumably, this error will be random and independent of the w_k sequence in the inertial error model. Furthermore, if external fixes are available only on an occasional basis, it is reasonable to assume that their errors will be uncorrelated. Thus v_k in the measurement model satisfies the necessary requirements, that is, it must also be a white sequence. With each position fix, we get a 2-tuple. Thus, the measurement model is

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & | & 0 & | & 0 \end{bmatrix}}_{\mathbf{H}_k} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_4 \\ x_5 \\ x_6 \\ \cdots \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad (6.1.10)$$

The \mathbf{H}_k matrix is indicated in Eq. (6.1.10), and \mathbf{R}_k will be assumed to be a diagonal matrix whose elements are the variances of the *N-S* and *E-W* components of external fix error. Since these will depend on the type of reference being used, the on-line computer must be programmed to accommodate to whatever references might be used. For example, one would not expect to obtain the same quality position fix using a handheld sextant as would be obtained from NAV-SAT or LORAN. The elements of \mathbf{R}_k must accurately reflect the quality of the measurement in order for the Kalman filter to produce good estimates. Once the estimates are obtained, the position and gyro bias corrections may be made in accordance with these estimates, and the inertial system is ready to operate unaided until another position fix becomes available.

The system will then be reset again and so forth. There are certain subtleties in applying corrections to the inertial system and the resulting mode of operation during the interim between fixes. Discussion of this aspect of the problem will be continued in Section 6.4.

Before leaving this example, it is especially important to note the versatility of the Kalman filter in accommodating measurements from a variety of sources and at arbitrary times. This is a dynamical problem. The mathematical connections among the nine quantities being estimated and the measurements are relatively complicated and are changing with time. In effect, the Kalman filter automatically accounts for these complex relationships via the key parameters ϕ_k , Q_k , \mathbf{H}_k , and \mathbf{R}_k ; then it makes the best use of whatever reference information that may be available, be it meager or plentiful, high in quality or low in quality! ■

6.2 ALTERNATIVE FORM OF THE DISCRETE KALMAN FILTER

The Kalman filter equations given in Chapter 5 can be algebraically manipulated into a variety of forms (4.5). An alternative form that is especially useful will now be presented. We begin with the expression for updating the error covariance, Eq. (5.4.26), and we temporarily omit the subscripts to save writing.

$$\mathbf{P} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^- \quad (6.2.1)$$

Recall that the Kalman gain is given by Eq. (5.4.21).

$$\mathbf{K} = \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1} \quad (6.2.2)$$

Substituting Eq. (6.2.2) into (6.2.1) yields

$$\mathbf{P} = \mathbf{P}^- - \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^- \quad (6.2.3)$$

We now wish to show that if the inverses of \mathbf{P} , \mathbf{P}^- , and \mathbf{R} exist, \mathbf{P}^{-1} can be written as

$$\mathbf{P}^{-1} = (\mathbf{P}^-)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \quad (6.2.4)$$

Justification of Eq. (6.2.4) is straightforward. We simply form the product of the right sides of Eqs. (6.2.3) and (6.2.4) and show that this reduces to the identity matrix. Proceeding as indicated,

$$\begin{aligned} & [\mathbf{P}^- - \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^-][(\mathbf{P}^-)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}] \\ &= \mathbf{I} - \mathbf{P}^-\mathbf{H}^T[(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1} - \mathbf{R}^{-1} + (\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^-\mathbf{H}^T\mathbf{R}^{-1}]\mathbf{H} \\ &= \mathbf{I} - \mathbf{P}^-\mathbf{H}^T[(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{I} + \mathbf{H}\mathbf{P}^-\mathbf{H}^T\mathbf{R}^{-1}) - \mathbf{R}^{-1}]\mathbf{H} \\ &= \mathbf{I} - \mathbf{P}^-\mathbf{H}^T[\mathbf{R}^{-1} - \mathbf{R}^{-1}]\mathbf{H} \\ &= \mathbf{I} \end{aligned}$$

An alternative expression for the Kalman gain may also be derived. Beginning with Eq. (6.2.2) we have

$$\mathbf{K} = \mathbf{P}^{-1} \mathbf{H}^T (\mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T + \mathbf{R})^{-1}$$

Insertion of $\mathbf{P}\mathbf{P}^{-1}$ and $\mathbf{R}^{-1}\mathbf{R}$ will not alter the gain. Thus, \mathbf{K} can be written as

$$\begin{aligned} \mathbf{K} &= \mathbf{P}\mathbf{P}^{-1} \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{R} (\mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T + \mathbf{R})^{-1} \\ &= \mathbf{P}\mathbf{P}^{-1} \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T + \mathbf{R}^{-1} + \mathbf{I})^{-1} \end{aligned}$$

We now use Eq. (6.2.4) for \mathbf{P}^{-1} and obtain

$$\begin{aligned} \mathbf{K} &= \mathbf{P}[(\mathbf{P}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1} + \mathbf{I})^{-1}] \\ &= \mathbf{P}(\mathbf{I} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{P}^{-1}) \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1} + \mathbf{I})^{-1} \\ &= \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{I} + \mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1}) (\mathbf{H} \mathbf{P}^{-1} \mathbf{H}^T \mathbf{R}^{-1} + \mathbf{I})^{-1} \\ &= \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} \quad (6.2.5) \end{aligned}$$

The main results have now been derived, and Eqs. (6.2.4) and (6.2.5) may be rewritten with the subscripts reinserted:

$$\mathbf{P}_k^{-1} = (\mathbf{P}_k^-)^{-1} + \mathbf{H}_k^T \mathbf{R}_k^{-1} \mathbf{H}_k \quad (6.2.6)$$

$$\mathbf{K}_k = \mathbf{P}_k \mathbf{H}_k^T \mathbf{R}_k^{-1} \quad (6.2.7)$$

Note that the updated error covariance can be computed without first finding the gain. Also, the expression for gain now involves \mathbf{P}_k ; therefore, if Eq. (6.2.7) is to be used, \mathbf{K}_k must be computed after the \mathbf{P}_k computation. Thus, the order in which the \mathbf{P}_k and \mathbf{K}_k computations appear in the recursive algorithm is reversed from that presented in Chapter 5. The alternative Kalman filter algorithm just derived is summarized in Fig. 6.1.

Note from Fig. 6.1 that two $(n \times n)$ matrix inversions are required for each recursive loop. If the order of the state vector is large, this leads to obvious computational problems. Nevertheless, the alternative algorithm has some useful applications. One of these will now be presented as an example.

Example 6.2 Suppose we wish to estimate an unknown constant based on a sequence of independent noisy measurements of the constant. We can think of the constant as being a deterministic random process that satisfies the differential equation

$$\dot{x} = 0 \quad (6.2.8)$$

Thus ϕ_k and Q_k are 1 and 0, respectively. Let us also speculate that very little is known about the process initially. The constant is equally likely to be positive or negative, and its magnitude could be quite large. It might be thought of as a random variable with a flat probability density function extending from $-\infty$ to $+\infty$. This being the case, the a priori estimate and associated error covariance should be

$$\hat{x}_0^- = 0 \quad (6.2.9)$$

$$P_0^- = \infty \quad (6.2.10)$$

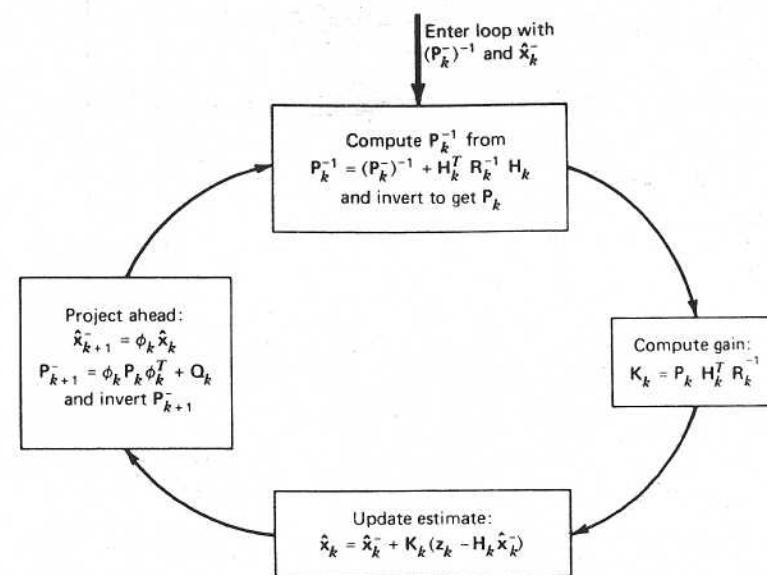


Figure 6.1. Alternative Kalman filter recursive loop.

However, this is not permitted in the usual Kalman filter algorithm (Fig. 5.7), because it leads to the indeterminant form ∞/∞ in the gain expression. The alternative algorithm will accommodate this situation, though, because $(P_0^-)^{-1}$ rather than P_0^- appears in the first step.

Time is of no consequence in this example because we are estimating a constant. So, let us assume we have N independent noisy measurements of x , each made at $t = 0$ and having an error variance of σ^2 . The measurement model is then

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} [x] + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} \quad (6.2.11)$$

and \mathbf{R}_k is the $(N \times N)$ diagonal matrix

$$\mathbf{R}_k = \begin{bmatrix} \sigma^2 & & & & \\ 0 & \sigma^2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I} \quad (6.2.12)$$

Proceeding with the first step of the alternative algorithm yields

$$\begin{aligned} P_0^{-1} &= (P_0^-)^{-1} + \mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0 \\ &= (\infty)^{-1} + [1 \ 1 \ \dots \ 1] \frac{1}{\sigma^2} \mathbf{I} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad (6.2.13) \end{aligned}$$

or

$$P_0 = \frac{\sigma^2}{N} \quad (6.2.14)$$

Next, the gain is computed as

$$\begin{aligned} \mathbf{K}_0 &= P_0 \mathbf{H}_0^T \mathbf{R}_0^{-1} \\ &= \frac{\sigma^2}{N} [1 \ 1 \ \dots \ 1] \begin{bmatrix} \frac{1}{\sigma^2} & 0 & \dots \\ 0 & \frac{1}{\sigma^2} & \\ \vdots & & \ddots \end{bmatrix} \\ &= \left[\frac{1}{N} \ \frac{1}{N} \ \dots \ \frac{1}{N} \right] \quad (6.2.15) \end{aligned}$$

Finally, the estimate is given by

$$\begin{aligned} \hat{x}_0 &= \hat{x}_0^- + \mathbf{K}_0(\mathbf{z} - \mathbf{H}_0 \hat{x}_0^-) \\ &= \mathbf{K}_0 \mathbf{z} = \frac{z_1}{N} + \frac{z_2}{N} + \dots + \frac{z_N}{N} \quad (6.2.16) \end{aligned}$$

The final result is no surprise; it is exactly the result one would expect from elementary statistics. The main point of this example is this: The alternative algorithm provides a means of starting the Kalman filter with "infinite uncertainty" if the physical situation under consideration so dictates. ■

6.3 SEQUENTIAL PROCESSING OF THE MEASUREMENT DATA

We now have two different Kalman filter algorithms as summarized in Figs. 5.7 and 6.1. They are, of course, algebraically equivalent and produce identical

estimates. The choice as to which should be used in a particular application is a matter of computational convenience. Both algorithms involve matrix inverse operations and these may lead to difficulties. When using the alternative algorithm of Fig. 6.1, there is no reasonable way to avoid two ($n \times n$) matrix inversions with each recursive cycle. If the dimension of the state vector n is large, this is, at best, awkward computationally. On the other hand, the matrix inverse that appears in the regular algorithm given in Fig. 5.7 is the same order as the measurement vector. Since this is often less than the order of the state vector, it is usually the preferred algorithm. Furthermore, if the measurement errors at time t_k are uncorrelated, the inverse operation can be eliminated entirely by processing the data sequentially. This will now be shown.

We begin with the expression for the updated error covariance, Eq. (6.2.6).

$$\mathbf{P}_k^{-1} = (\mathbf{P}_k^-)^{-1} + [\mathbf{H}_k^{aT} \mid \mathbf{H}_k^{bT} \mid \dots] \begin{bmatrix} (\mathbf{R}_k^a)^{-1} & 0 & 0 \\ 0 & (\mathbf{R}_k^b)^{-1} & 0 \\ \vdots & \vdots & \ddots \\ 0 & 0 & \ddots \end{bmatrix} \begin{bmatrix} \mathbf{H}_k^a \\ \mathbf{H}_k^b \\ \vdots \end{bmatrix} \quad (6.3.1)$$

The second term in Eq. (6.3.1) is intentionally written in partitioned form and \mathbf{R}_k is assumed to be at least block diagonal. Physically, this means that the measurements available at t_k can be grouped together such that the measurement errors among the a, b, \dots blocks are uncorrelated. This is often the case when redundant measurements come from different instruments. We next expand the partitions of Eq. (6.3.1) to get

$$\mathbf{P}_k^{-1} = (\mathbf{P}_k^-)^{-1} + \underbrace{\mathbf{H}_k^{aT} (\mathbf{R}_k^a)^{-1} \mathbf{H}_k^a}_{\mathbf{P}_k^{-1} \text{ after assimilating block } a \text{ measurements}} + \mathbf{H}_k^{bT} (\mathbf{R}_k^b)^{-1} \mathbf{H}_k^b + \dots \quad (6.3.2)$$

\mathbf{P}_k^{-1} after assimilating both block a and b measurements

and so forth

Note that the sum of the first two terms is just the \mathbf{P}_k^- one would obtain after assimilating the "block a " measurement just as if no further measurements were available. The Kalman gain associated with this block of measurements may now be used to update the state estimate accordingly. Now think of making a trivial projection ahead through zero time. The a posteriori \mathbf{P} then becomes the a priori \mathbf{P} for the next step. When this is added to the b term of Eq. (6.3.2), we have the updated \mathbf{P}_k^{-1} after assimilating the second block of data. This can now be repeated sequentially until all blocks are processed. The final estimate and associated error is then the same as would be obtained if all the measurements at t_k had been processed simultaneously. Thus, the designer has

some flexibility in the design of the system software. The available measurements at any particular time may be processed either sequentially in blocks, or all at once, as best suits the situation at hand.

Sequential processing is also useful from a system organization viewpoint. Often the system must have the flexibility to accommodate a variety of measurement combinations at each update point. By block processing, the system may be programmed to cycle through all possible measurement blocks sequentially, processing those that are available and skipping those that are not. Simultaneous processing requires a somewhat more complicated system organization whereby the system must be able to form appropriate \mathbf{H}_k and \mathbf{R}_k matrices for all possible combinations of measurements, and it must be prepared to do the corresponding matrix operations with various dimensionality.

There are bound to be some applications where the measurement errors are all mutually correlated. The \mathbf{R}_k matrix is then "full." If this is the case, and sequential processing is desirable, linear combinations of the measurements may be formed in such a way as to form a new set of measurements whose errors are uncorrelated. One technique for accomplishing this is known as the Gram-Schmidt orthogonalization procedure (6). This procedure is straightforward and an exercise is included to demonstrate its application to the problem of decoupling the measurement errors (see Problem 6.4).

6.4 AIDED INERTIAL NAVIGATION SYSTEMS

Kalman filtering has been used extensively in navigation applications since the mid-1960s, and interest continues unabated with the most recent applications having to do with the new navigation satellite system known as Global Positioning System (GPS) (7,8). Navigation problems seem to form a natural setting for Kalman filtering for the following reasons:

1. The dynamics are usually linear (or may be linearized with reasonable accuracy—see Section 9.1).
2. There is often a redundancy of measurement information from a variety of navigational sources.
3. The navigation problem is basically an on-line problem. Measurements must be processed essentially in real time, and thus efficient processing is a necessity.
4. There is frequently a need to squeeze the best possible performance out of the system, and thus the data processing must be optimal (or at least nearly so).

One particular integrated navigation system problem that keeps recurring is that of integrating inertial system measurements with other navigation aids such as LORAN or NAV SAT-data. It was mentioned in Section 6.1 that

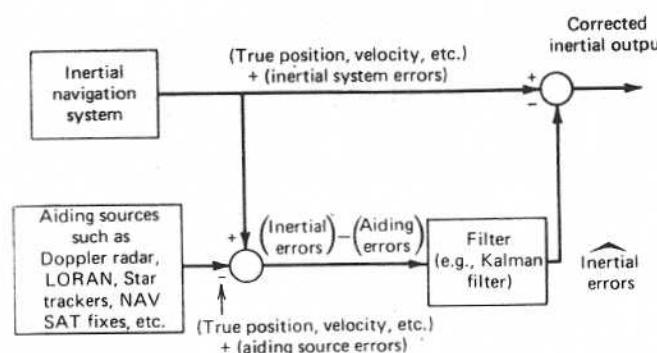


Figure 6.2. Aided inertial navigation system.

inertial navigation systems have drift characteristics that cause the system errors to grow with time. They also exhibit undamped oscillatory errors that are undesirable (2,3). Thus, in applications where the mission time is relatively long, inertial systems are usually used in conjunction with other navigation instruments which perform an aiding role in the overall navigation system. Such integrated systems are sometimes called *aided* inertial systems, in contrast to *pure* inertial systems, which operate without the benefit of any supplementary navigation information.

Since aided inertial systems always involve redundant measurements, the question arises as to how all the redundant information should be blended together to yield the best estimates of position, velocity, and so on. The system configuration usually used is shown in Fig. 6.2.* This will be recognized as being similar to the complementary filter discussed in Section 4.6. Philosophically, we think of the inertial system as being the prime system that provides continuous outputs of position, velocity, and vehicle angular orientation relative to some known reference such as an earth-fixed reference frame. The aiding sources are supplementary to the inertial system and their outputs may only be available intermittently, depending on the situation at hand.

The purpose of the filter is to estimate the inertial system errors. The estimates may then be used to correct the inertial system's output as indicated in Fig. 6.2. It is important to note that the filter, be it optimal or otherwise, operates only on the system errors. The total navigation quantities, such as position and velocity, pass through the system without distortion or delay. This

*The configuration shown in Fig. 6.2 is known as the feedforward configuration. It should be apparent that the estimates of the inertial errors could also be fed back and the corrections made internally in the inertial system. This is called the feedback configuration. Within the linearity assumption, both configurations yield the same results. However, in practice there are subtle differences. References 9, 10, and 11 may be consulted for further discussion of aided inertial systems and the relative merits of the feedforward and feedback configurations.

is important in the navigation problem, because vehicle position and velocity are usually *not* properly modeled as random processes. If the filter indicated in Fig. 6.2. is to be a Kalman filter, then all the errors in the inertial and aiding sources must be modeled in vector form, and the filter must estimate all the state variables of the system. Another navigation example will illustrate the versatility of the Kalman filter in this mode of operation.

Example 6.3. The system integration of this example was originally proposed in the early 1960s and it represents one of the early attempts to achieve a fully integrated navigation system using Kalman filter techniques (12). Even though the system was never produced in quantity (primarily because of cost), it still serves as an excellent system-integration example because of the particular blend of available measurements. The system organization is as shown in Fig. 6.2, and the sources of information available for aiding the marine inertial system of this example are as follows:

1. Ship's velocity relative to the water. This comes from an electromagnetic log which is quite accurate. However, uncertainties in ocean currents degrade the accuracy as an absolute velocity measurement.
2. Occasional discrete position fixes. These might come from various sources such as LORAN or other navigation aids. Discrete position fixes are assumed to be available only occasionally and not on any fixed time schedule.
3. Single-body celestial angle measurements from a radiometric sun/moon tracker. This device is capable of tracking the sun's electromagnetic radiation in much the same manner that a star tracker tracks a star's optical radiation. It can track through cloud cover, though, which provides all-weather capability. The accuracy is quite good while tracking the sun. There is some loss of accuracy though while tracking the moon, because it is a weaker radio source. Therefore, the measurement error variance is larger when tracking the moon than when tracking the sun. Obviously, these measurements are only available when the sun (or moon) is above the horizon.

We now proceed to model the error sources in the system. The undamped inertial error equations for a slow-moving vehicle are*

$$\begin{aligned}\dot{\psi}_x - \Omega_z \psi_y &= \varepsilon_x \\ \dot{\psi}_y + \Omega_z \psi_x - \Omega_x \psi_z &= \varepsilon_y \\ \dot{\psi}_z + \Omega_x \psi_y &= \varepsilon_z \\ \ddot{\theta}_x + \omega_0^2 \delta \theta_x + \omega_0^2 \psi_x - 2\Omega_z \delta \theta_y &= -\delta a_x/R \\ \ddot{\theta}_y + \omega_0^2 \delta \theta_y + \omega_0^2 \psi_y + 2\Omega_z \delta \theta_x &= \delta a_y/R\end{aligned}\quad (6.4.1)$$

*Just as in Example 6.1, the reader who is not familiar with inertial system error analysis can simply accept Eqs. (6.4.1) as accurately describing the inertial error propagation and proceed on with the Kalman filter modeling. Appropriate references are given in Section 6.1.

where

ψ_x, ψ_y, ψ_z = Linear combinations of angular platform and position errors (in radians). (Certain components of ψ are sensed by the radiometric tracker as explained later. Note that ψ_x and ψ_y are not simply position errors in this error model.)

$\delta \theta_x, \delta \theta_y$ = East and North position errors (in radians).

$\delta \theta_x, \delta \theta_y$ = East and North velocity errors (in rad/sec).

Ω_x, Ω_z = Components of earth rate on North and Up platform axes (assumed known).

$$\omega_0^2 = (\text{Earth's gravity constant})/(\text{Earth radius}) = g/R \left(\frac{\text{rad}}{\text{sec}} \right)^2$$

$\varepsilon_x, \varepsilon_y, \varepsilon_z$ = Gyro biases (rad/sec).

$\delta a_x, \delta a_y$ = x and y accelerometer biases. (Distance units/sec²).

Note that this is a seventh-order linear dynamical system with the gyro and accelerometer instrument biases acting as the driving functions. The first seven states may be defined as: $x_1 = \psi_x, x_2 = \psi_y, x_3 = \psi_z, x_4 = \delta \theta_x, x_5 = \delta \theta_y, x_6 = \delta \theta_z, x_7 = \delta \theta_y$. Next, the gyro and accelerometer instrument errors look more like quasi-biases than white noises. Thus, these will be modeled as Markov processes and the state vector will be augmented with five more state variables. Thus let $x_8 = \varepsilon_x, x_9 = \varepsilon_y, x_{10} = \varepsilon_z, x_{11} = -\delta a_y/R$, and $x_{12} = \delta a_z/R$. These state variables satisfy first-order differential equations of the form

$$\dot{x}_8 + \beta_8 x_8 = \sqrt{2\sigma_8^2 \beta_8} f_8$$

$$\dot{x}_{12} + \beta_{12} x_{12} = \sqrt{2\sigma_{12}^2 \beta_{12}} f_{12} \quad (6.4.2)$$

and these must be appended to the original seven state equations. Thus, the inertial system error model requires a total of 12 state variables.

Further augmentation of the state vector is now required because of the errors of the aiding sources. The ocean-current uncertainties may be expected to contribute a sizable error to the aiding velocity measurement. Since these will be slowly varying random quantities, they will also be modeled as Markov processes. Thus, we have the x and y slowly varying components of velocity reference error modeled as

$$\dot{x}_{13} + \beta_{13} x_{13} = \sqrt{2\sigma_{13}^2 \beta_{13}} f_{13}$$

$$\dot{x}_{14} + \beta_{14} x_{14} = \sqrt{2\sigma_{14}^2 \beta_{14}} f_{14} \quad (6.4.3)$$

Also, since there are relatively slowly varying errors in the radiometric sun-tracker that cannot be ignored, these will also be modeled as Markov processes. The sun-tracker senses the angular deviation of the sun's line-of-sight from the boresight of the antenna. If the antenna is pointed in accordance with the known sun position and the inertial system output (with its associated position and platform errors), the measured quantities are the two components of ψ orthogonal to the line of sight (2). The long-time-constant errors associated with these measurements will also be modeled as Markov processes and will be denoted as x_{15} and x_{16} . This brings the total number of state variables to 16. The discrete position-fix errors will be assumed to be uncorrelated so that no further augmentation is needed for them. The final state model is summarized in Fig. 6.3.

The measurement model for the discrete position fixes is straightforward. After differencing with the corresponding inertial latitude and longitude, these measurements become direct one-to-one samples of state variables x_4 and x_6 . They will be denoted as z_1 and z_2 and we will assume the measurement errors are uncorrelated and white with variances r_{11} and r_{22} .

It will be assumed that the ship's heading is known with sufficient accuracy to resolve the EM log velocity measurement into East and North components. These will be assumed to be sampled (with perhaps some prefiltering) and compared with the corresponding inertial velocities, and the resulting differences will be denoted as measurements z_3 and z_4 . According to our model, they will contain Markov components as well as the respective inertial errors. In addition, we will allow for small additive white components due to instrument errors. The white errors in z_3 and z_4 will be assumed to be uncorrelated and to have variances r_{33} and r_{44} .

The radiometric tracker provides a measure of the two components of ψ normal to the line-of-sight to the celestial body being tracked. These, in turn, are related to ψ_x , ψ_y , ψ_z via the direction cosines relating the x, y, z platform axes to the antenna orthogonal axes, say u and v . We denote these direction cosines as C with appropriate subscripts and simply note that they can be computed on-line, given the ship's approximate position and the ephemeris data for the celestial body being tracked. Note that they will vary with time, and they must be recalculated with each measurement sample. It should also be mentioned that in this mode of operation the "tracking" loops are closed via the corrections that are applied to the inertial system. In the interest of keeping the ψ errors small, the system is envisioned as operating in the feed back configuration rather than the feed forward configuration shown in Fig. 6.2. Just as in the case of the velocity measurements, the tracker measurement errors are assumed to be an additive combination of white and Markov components.

The final measurement model is given by Eq. (6.4.5). Recall that the six-direction cosines in the lower left corner of the \mathbf{H} matrix are functions of time and must be computed on-line. Similarly, the Ω_x and Ω_z terms in the \mathbf{F} matrix of Fig. 6.3 vary slowly with the ship's latitude. Thus the state transition matrix must be occasionally updated, even though the sampling interval is held con-

$$\begin{aligned}
 & \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \\ x_{16} \end{array} \right] = \\
 & \left[\begin{array}{c} 0 \\ -\Omega_x \\ 0 \\ -\Omega_z \\ 0 \\ -\omega_0^2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] + \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{g_8}^2\beta_8} f_8 \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{g_9}^2\beta_9} f_9 \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_0}^2\beta_{10}} f_{10} \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_1}^2\beta_{11}} f_{11} \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_2}^2\beta_{12}} f_{12} \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_3}^2\beta_{13}} f_{13} \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_4}^2\beta_{14}} f_{14} \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_5}^2\beta_{15}} f_{15} \\
 & \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right] \sqrt{2\sigma_{i_6}^2\beta_{16}} f_{16} \\
 \end{aligned} \tag{6.4.4}$$

Figure 6.3. State model for Example 6.3.

The six measurement quantities in this example are logically grouped in pairs and may be processed sequentially in that manner. Then, if any particular pair is not available at the scheduled update time, it is simply skipped. (For instance, both the sun and moon might be below the horizon.) This is a good example of the benefits of sequential processing.

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ C_{ux} & C_{uy} & C_{uz} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ C_{vx} & C_{vy} & C_{vz} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix} \quad (6.4.5)$$

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad x_8 \quad x_9 \quad x_{10} \quad x_{11} \quad x_{12} \quad x_{13} \quad x_{14} \quad x_{15} \quad x_{16}$

Fortunately, in this example, the natural periods of oscillation are relatively large. The ψ -equation oscillatory period is 24 hours, and the level loops have a period of about 84 minutes (the so-called Schuler period). Thus, if the update cycle time is set at about two minutes, this provides an adequate sampling rate as well as allowing ample time for computer calculations between updates.

Figure 6.4 shows an rms error plot for a typical situation where the only sources aiding the inertial system are external reference velocity and sun-tracking. The ship was assumed to be at 45 deg latitude, the sun at equinox, and the active tracking period was from an hour after sunrise to an hour before sunset. The Kalman filter program was allowed to run long enough to reach a stable steady-state condition, and the result is shown in Fig. 6.4. The various system parameters were chosen to be typical of a modest accuracy system, and their actual values are not especially important for this discussion. The relative results are of interest though and are much as one would expect:

1. Longitude error is smallest after sunrise and just before sunset.
2. Latitude error is minimum around noon when the sun is in the south.
3. There is a relatively large drift at night when celestial observations are not available.

It is important to note that the error plots of Fig. 6.4 are the result of covariance analysis and not simulation using random numbers. That is, the two

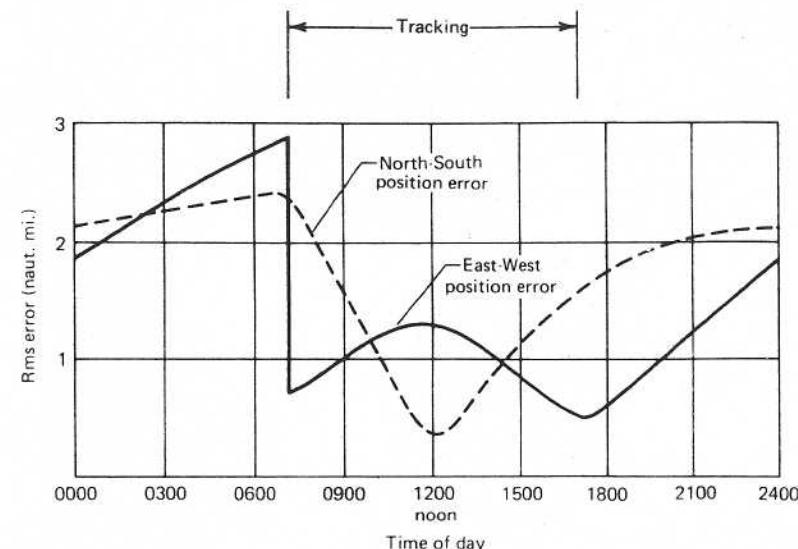


Figure 6.4. Rms position errors for Example 6.3.

plots are simply the square roots of the p_{44} and p_{66} terms of the \mathbf{P} matrix after the steady-state condition (with a 24-hour period) was reached. These results then represent the rms errors one would expect under ideal conditions where the Kalman filter model faithfully represents the actual situation at hand. It is interesting to note that a stable condition was reached using a Kalman filter mode of operation with only reference velocity and sun-tracking as the aiding sources. This may seem to be obvious from the outset. However, this was not so in the early days of the system analysis because of certain subtleties regarding the observability of the system. This will be continued in Section 6.6.

The lesson to be learned from this example is simply this. The system optimization problem presented by the particular mix of measurement data in this application is hopelessly complicated when viewed from the Wiener or batch-processing viewpoint. On the other hand, the system organization for recursively processing the measurement data is straightforward once the block diagram of Fig. 6.2 and the format requirements of a Kalman filter are understood. The dimensionality of the filter may grow with additional aiding sources and refinements in the state model. However, the methodology remains the same, and this is the beauty of the Kalman filter solution of the system integration problem. (See Section 6.6 for further comments on modeling.) ■

6.5 POWER SYSTEM RELAYING APPLICATION

New applications of Kalman filtering keep appearing regularly, and many of these are now outside the original application area of navigation. One such

recent application has to do with power system relaying (13,14). When a fault (short) occurs on a three-phase transmission line, it is desirable to sense the problem promptly and take appropriate relaying action to protect the remainder of the system. The hierarchy of decisions that must be made as to which relays should trip (and where) is relatively complicated. It suffices to say here that it is desirable to determine the distance to the fault as soon as possible; and, in order to do this, the steady-state postfault currents and voltages must be estimated. Transient components are superimposed on the steady-state signals immediately after the fault, so that the transients become the corrupting noise in the problem. Normally, these transients are not considered as random noise. However, to model them otherwise, complicates the problem immensely because of the many variables involved. So, the basic problem is to estimate the steady-state components of the sending-end voltages and currents in the presence of the transient (noise) components. It is assumed that digital samples of the various phase voltages and currents are available for processing at a reasonably fast rate, say 64 samples per cycle of the 60-Hz signal.

Girgis and Brown (14) made an extensive simulation study of the transients accompanying various types of faults for various lengths of line, and so forth. They concluded that the transients could be approximated as nonstationary random processes, and they developed models accordingly that would fit the required format of a Kalman filter. The simulation study indicated that the voltage transients consisted primarily of high-frequency components that decayed exponentially with time. Thus it seemed reasonable to model the time samples of this process as a white sequence with an exponentially decaying variance. The current transients, however, showed (on the average) a sizable long-time-constant exponential component in addition to the high-frequency components. (Power engineers sometimes refer to this as the "dc offset.") Thus the model chosen for the current noise was an exponential process with random initial amplitude plus a white sequence with exponentially decaying variance. The signal process to be estimated for both current and voltage was a sine wave with random amplitude and phase. This is readily modeled as a two-element vector, where the state variables are the coefficients of the sine and cosine components of the wave [see Section 5.2 and Problem 5.7(b)]. The final models for the currents and voltages may be summarized as follows:

Voltage Model (same for each phase)

(1) State equations:

$$\begin{bmatrix} x1_{k+1} \\ x2_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x1_k \\ x2_k \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (6.5.1)$$

(2) Measurement equation:

$$z_k = [\cos \omega_0 k \Delta t - \sin \omega_0 k \Delta t] \begin{bmatrix} x1_k \\ x2_k \end{bmatrix} + v_k \quad (6.5.2)$$

(3) Initial conditions:

$$\mathbf{P}_0^- = \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix} \quad (6.5.3)$$

$$\hat{\mathbf{x}}_0^- = \text{Measured value of } \mathbf{x} \text{ just prior to the fault} \quad (6.5.4)$$

The σ_v^2 parameter was determined by simulation. It is fixed and not determined on-line. On the other hand, the initial estimates of \mathbf{x} are assumed to be determined on-line. There is no reason to waste the available measurement information just prior to the fault. It should be obvious from Eq. (6.5.1) that $\mathbf{Q}_k = 0$ for the voltage model. Also, as mentioned previously, \mathbf{R}_k is assumed to decay exponentially with k , and the exponential parameters are predetermined by simulation or experimental data.

Current Model (same for all phases)

(1) State equations:

$$\begin{bmatrix} x1_{k+1} \\ x2_{k+1} \\ x3_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-\beta \Delta t} \end{bmatrix} \begin{bmatrix} x1_k \\ x2_k \\ x3_k \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w_k \end{bmatrix} \quad (6.5.5)$$

(2) Measurement equation:

$$z_k = [\cos \omega_0 k \Delta t - \sin \omega_0 k \Delta t] \begin{bmatrix} x1_k \\ x2_k \\ x3_k \end{bmatrix} + v_k \quad (6.5.6)$$

(3) Initial conditions:

$$\mathbf{P}_0^- = \begin{bmatrix} \sigma_i^2 & 0 & 0 \\ 0 & \sigma_i^2 & 0 \\ 0 & 0 & \sigma_i^2 \end{bmatrix} \quad (6.5.7)$$

$$\hat{\mathbf{x}}_0^- = \begin{bmatrix} x1^- \text{ (meas.)} \\ x2^- \text{ (meas.)} \\ 0 \end{bmatrix} \quad (6.5.8)$$

Just as in the voltage model, σ_i^2 is determined off-line and the first two elements of $\hat{\mathbf{x}}_0^-$ are obtained from measurements just prior to the fault. The third element of $\hat{\mathbf{x}}_0^-$ (i.e., the exponential component) is assumed to be zero. The measurement error variance R_k is assumed to decay exponentially just as in the voltage model. The \mathbf{Q}_k parameter is not zero, though, because the exponential component was observed in the simulation studies to have a small residual noise

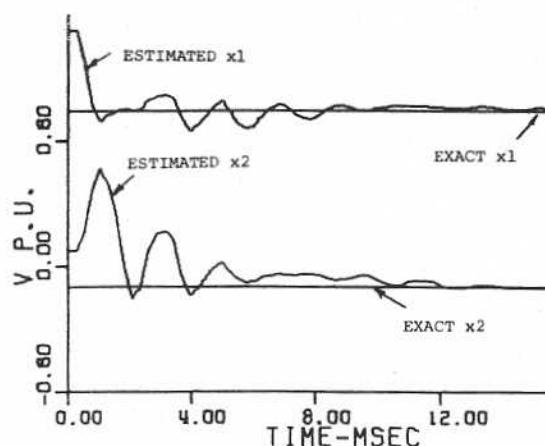


Figure 6.5. Kalman filter estimation of the postfault voltage states.

associated with it. This is accounted for in the model with w_k , and thus the "33" element of Q_k is nonzero. In effect, x_3 is modeled as a nonstationary process with a large random initial value, and then it relaxes to a Markov process with a relatively small rms value in the steady-state condition. This is an unusual model, but perfectly legitimate. Again, this illustrates the versatility of the Kalman filter to adapt to a wide variety of situations.

Figures 6.5 and 6.6 show the voltage and current estimates for a particular simulation of a line-to-ground fault located 90 miles from the sending end. The details of the simulation are not important here, because the results are all relative. Recall that x_1 and x_2 are the coefficients of the sine and cosine components of the steady-state values, so they are constants. Note the Kalman filter estimates converge reasonably well to the correct values after about 8 ms (half cycle at 60 Hz). Figure 6.7 shows the result of using the voltage and current estimates of this simulation to compute distance to the fault. A similar distance calculation was also made using currents and voltages as determined by a discrete Fourier transform algorithm. Both the Fourier transform and Kalman

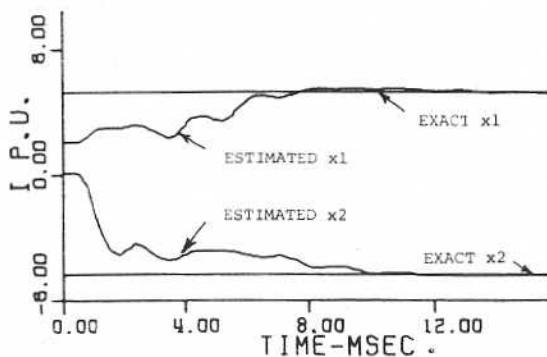


Figure 6.6. Kalman filter estimation of the postfault current states.

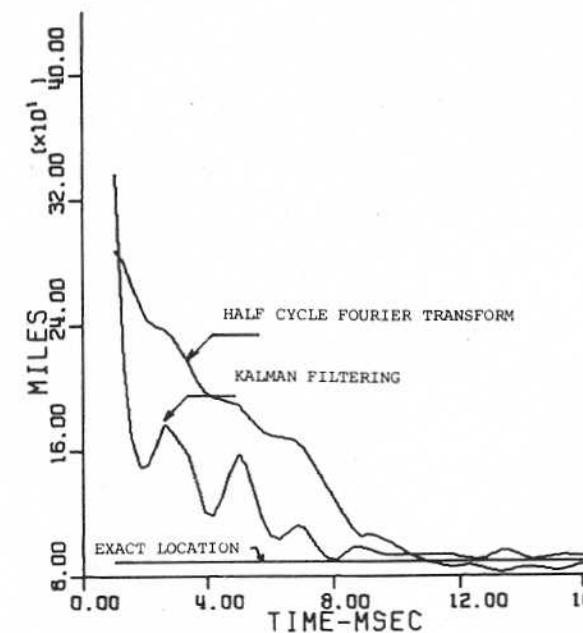


Figure 6.7. The computed distance to the fault using the Kalman filter algorithm and the discrete Fourier transform.

filter results are shown in Fig. 6.7, and it is clear that the Kalman filter converges on the correct result faster than the other algorithm. This is as expected. The discrete Fourier transform approach does not account for the time-varying nature of the noise, nor does it allow for any a priori knowledge of the parameters being estimated. Of course, both algorithms converge to the correct result eventually. However, time is of the essence! Why accept inferior performance when optimal performance is readily available.

6.6 DIVERGENCE PROBLEMS

Since the discrete Kalman filter is recursive, the looping can be continued indefinitely, in principle at least. There are practical limits, though, and under certain conditions divergence problems can arise. We elaborate briefly on three common sources of difficulty.

Roundoff Errors

As with any numerical procedure, roundoff error can lead to problems as the number of steps becomes large. This is especially true in on-line applications

where computer constraints sometime dictate fixed-point arithmetic. This always leads to difficulty when the dynamic range of the variables is large. There is no one simple solution to the problem, and each case has to be examined on its own merits. Fortunately, if the system is observable, the Kalman filter has a degree of natural stability. In this case a stable, steady-state solution for the \mathbf{P} matrix will normally exist, even if the process is nonstationary. If the \mathbf{P} matrix is perturbed from its steady-state solution in such a way as to not lose positive definiteness, then it tends to return to the same steady-state solution. This is obviously helpful, provided \mathbf{P} does not lose its positive definiteness.

Some techniques that have been found useful in preventing, or at least forestalling, roundoff error problems are:

1. Avoid fixed-point arithmetic, if at all possible. Do not hesitate to use double-precision arithmetic if in doubt, especially in off-line analysis work.
2. If measurement data is sparse, beware of propagating the \mathbf{P} matrix in many tiny steps between measurements. (This simplifies the transition matrix calculation but opens "Pandora's box" with regard to roundoff error, unless higher-order precision arithmetic is used.)
3. If possible, avoid deterministic (undriven) processes in the filter modeling. (*Example*: a random constant.) These usually lead to a situation where the \mathbf{P} matrix approaches a semi-definite condition as the number of steps becomes large. A small error may then trip the \mathbf{P} matrix into a non-positive-definite condition, and this can then lead to divergence. A good solution is to add (if necessary) small positive quantities to the major-diagonal terms of the \mathbf{Q} matrix. This amounts to inserting a small amount of process noise to each of the states. This leads to a degree of suboptimality, but that is better than having the filter diverge!
4. Symmetrize the \mathbf{P} and \mathbf{P}^{-} matrices with each recursive step. We know a covariance matrix must be symmetric, so any asymmetry must be due to roundoff error. The asymmetry can grow if left unchecked. (The symmetry problem is automatically solved if, in programming the recursive equations, one assumes symmetry and uses only the upper (or lower) triangular part of the covariance matrix in all required matrix-multiply operations.)
5. If all the usual precautions fail, there exists an algorithm for propagating the matrix "square root" of \mathbf{P} , rather than \mathbf{P} itself (see Section 9.3). This is helpful in avoiding extreme ranges in the terms of the \mathbf{P} matrix. (There are those who would say that one of the square root algorithms should *always* be used because of their superior numerical behavior. This may be, but it does mean extra effort in programming.)

Modeling Errors

Another type of divergence may arise because of inaccurate modeling of the process being estimated. (Also see Section 9.1.) This has nothing to do with numerical round-off; it occurs simply because the designer (engineer) "told"

the Kalman filter that the process behaved one way; whereas, in fact, it behaves another way. As a simple example, if you tell the filter that the process is a random constant (i.e., zero slope), and the actual process is a random ramp (nonzero slope), the filter will be continually trying to fit the wrong curve to the measurement data! This can also occur with nondeterministic as well as deterministic processes, as will now be demonstrated.

Example 6.4 Consider a process that is actually random walk but is incorrectly modeled as a random constant. We have then (with numerical values inserted to correspond to a subsequent simulation):

- (a) The "truth model":
 $\dot{x} = w(t)$, $w(t)$ = unity Gaussian white noise, and $\text{Var}[x(0)] = 1$
 $z_k = x_k + v_k$, measurement samples at $t = 0, 1, 2, \dots$ and $\text{Var}(v_k) = .1$.

- (b) Incorrect Kalman filter model:

$$x = \text{constant}, \quad \text{where } x \sim N(0,1)$$

$$z_k = x_k + v_k \quad (\text{same as for truth model})$$

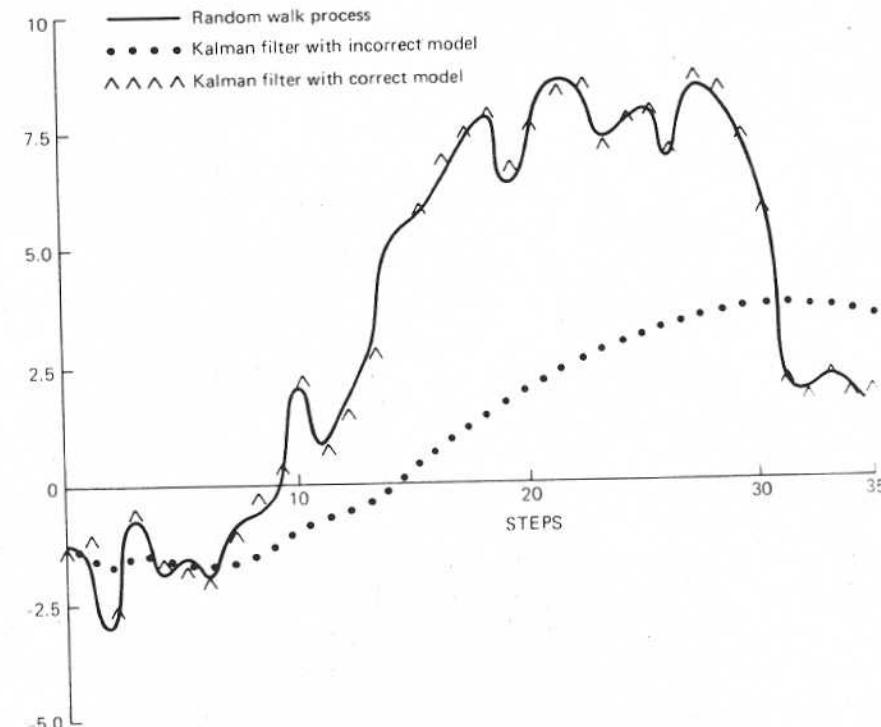


Figure 6.8. Simulation results for random walk example.

The Kalman filter parameters for the incorrect model (b) are: $\phi_k = 1$, $Q_k = 0$, $H_k = 1$, $R_k = .1$, $\dot{x}_0 = 0$, and $P_0^- = 1$. For the truth model the parameters are the same except that $Q_k = 1$, rather than zero.

The random walk process (a) was simulated using Gaussian random numbers with zero mean and unity variance. The resulting sample process for 35 sec is shown in Fig. 6.8. A measurement sequence z_k of this sample process was also generated using another set of $N(0, 1)$ random numbers for v_k . This measurement sequence was first processed using the incorrect model (i.e., $Q_k = 0$), and again with the correct model (i.e., $Q_k = 1$). The results are shown along with the sample process in Fig. 6.8. In this case, the measurement noise is relatively small ($\sigma \approx .3$), and we note that the estimates of the correctly modeled filter follow the random walk quite well. On the other hand, the incorrectly modeled filter does very poorly after the first few steps. This is due to the filter's gain becoming less and less with each succeeding step. At the 35th step the gain is almost two orders of magnitude less than at the beginning. Thus, the filter becomes very sluggish and will not follow the random walk. Had the simulation been allowed to go on further, it would have become even more sluggish. ■

The moral to Example 6.4 is simply this. Any model that assumes the process, or any facet of the process, to be absolutely constant forever and ever is a risky model. In the physical world, very few things remain absolutely constant. Instrument biases, even though called "biases," have a way of slowly changing with time. Thus, most instruments need occasional recalibration. The obvious remedy for this type of divergence problem is always to insert some process noise into each of the state variables. Do this even at the risk of some degree of suboptimality; it makes for a much safer filter than otherwise. It also helps with potential roundoff problems. (Note: Don't "blame" the filter for this kind of divergence problem. It is the fault of the designer/analyst, not the filter!)

In view of the preceding remarks, it is of interest to look again at the gyro-bias models used in Examples 6.1 and 6.3. In the Bona-Smay example the gyro biases were modeled as the sum of Markov and constant components. Obviously the constant component is somewhat risky if the system is intended to operate for long periods of time. It is completely inflexible, and the filter will give measurement data that is months old just as much weight as current data in estimating the constant. This can be corrected simply by changing the q_{77} , q_{88} , and q_{99} terms in the Q matrix from zero to appropriate small positive values. In effect, this allows the constants the flexibility to random walk if the measurements so dictate. The simple, one-state Markov model assumed in the radio-sextant/inertial example (Example 6.3) is also risky, but for a different reason. The Markov process is flexible in the sense that it can "wander around," but it is assumed to do this about a mean value of zero. Thus it will not adjust itself to

any major shift in bias that might occur during the course of the mission. This can be corrected with the addition of a random-walk component. The price, of course, is extra complexity.

Choosing an appropriate process model is always an important consideration, and a certain amount of common sense judgment is called for in deciding on a model that will fit the situation at hand reasonably well, but at the same time will not be too complicated. There are no firm rules on this. You have to consider each case on its own merits.

Observability Problem

There is a third kind of divergence problem that may occur when the system is not observable. Physically, this means that there are one or more state variables (or linear combinations thereof) that are hidden from the view of the observer (i.e., the measurements). As a result, if the unobserved processes are unstable, the corresponding estimation errors will be similarly unstable. This problem has nothing to do with roundoff error or inaccurate system modeling. It is just a simple fact of life that sometimes the measurement situation does not provide enough information to estimate all the state variables of the system. In a sense, the problem should not even be referred to as divergence, because the filter is still doing the best estimation possible under adverse circumstances.

There are formal tests of observability that may be applied to systems of low dimensionality. These tests are not always practical to apply, though, in higher-order systems. Sometimes one is not even aware that a problem exists until after extensive error analysis of the system (see Section 6.7). If unstable estimation errors exist, this will be evidenced by one or more terms along the major diagonal of \mathbf{P} tending to increase without bound. If this is observed, and proper precautions against roundoff error have been taken, the analyst knows an observability problem exists. We will now return to the integrated navigation system example of Section 6.4 to illustrate this phenomenon.

Development of a radiometric sun-tracker as an all-weather navigation aid began in the early 1950s (15). This preceded Kalman's recursive filtering work by about a decade, so methods available for system integration at that time were relatively simple. Of course, some sort of inertially stabilized vertical reference was needed for the radio sextant. It was also generally recognized that the inertial reference (equipped with accelerometers as well as gyros) could function as a navigation device in its own right and provide position, velocity, and heading information as well as a vertical reference. In the early years of development, the system integration was more or less taken for granted. If nothing else, it was assumed that the radio sextant could provide lines of position at appropriate times during the day, these could then be extrapolated ahead using inertially derived velocity, and a sequence of position fixes would be obtained in much the same manner as navigators have been

doing for centuries (16). In the early 1960s, this mode of operation was shown to be unworkable (12). This can best be seen by referring to the “ ψ equations” [first three equations of Eq. (6.4.1)] given in Example 6.3. Recall that only the two components of ψ normal to the line-of-sight can be observed with the radio sextant. The component along the line-of-sight cannot be observed. If the three ψ equations are referred to an inertially fixed reference frame rather than the earth-fixed xyz frame, an especially simple equation results*:

$$\left[\frac{d\psi}{dt} \right]_{\text{inertial}} = \epsilon \quad (6.6.1)$$

This says that with zero gyro drift (i.e., $\epsilon = 0$) the spacial vector ψ will remain constant in inertial space. But, the line-of-sight from the earth to a single celestial body also stays fixed in inertial space (except for a small parallax effect). Thus, the component of ψ along the line-of-sight can never be observed with just single-body observations. It, in turn, reflects into the position errors, so certain components of the inertial system position error cannot be corrected, based on single-body measurements alone. Thus, the proper method of system integration was not at all obvious after this discovery.

It works out that the external velocity reference information plays a key role in stabilizing the errors in this system. Without it, the errors grow without bound. The system is formally not observable if the two components of ψ normal to the line-of-sight are the only aiding measurements. When the two components of inertial velocity error are added to the measurement vector, the system then becomes observable. This is confirmed from Kalman filter error covariance runs for the two measurement situations. Typical latitude rms error plots are shown in the sketch of Fig. 6.9. The instability without reference velocity information is obvious.

This is a case study where Kalman filtering worked out to be the hero of the story. It came on the scene just in time to provide a good workable solution to the system integration problem. Without it, the external velocity information, with its accompanying error due to ocean current uncertainties, might have been incorporated into the system improperly, which would have resulted in a larger system error than necessary.

*This can be verified using a theorem from mechanics that is sometimes called the Coriolis theorem:

$$\left[\frac{d\mathbf{A}}{dt} \right]_{\substack{\text{Fixed} \\ \text{reference} \\ \text{frame}}} = \left[\frac{d\mathbf{A}}{dt} \right]_{\substack{\text{Rotating} \\ \text{reference} \\ \text{frame}}} + \boldsymbol{\omega} \times \mathbf{A}$$

In this equation, \mathbf{A} is any spacial vector, $\boldsymbol{\omega}$ is the angular rate of rotation of the rotating reference frame, and the indicated differentiations are performed on the components of \mathbf{A} in the respective reference frames. The ψ equations in the earth-fixed (rotating) reference frame are now obtained by applying the Coriolis theorem to Eq. (6.6.1).

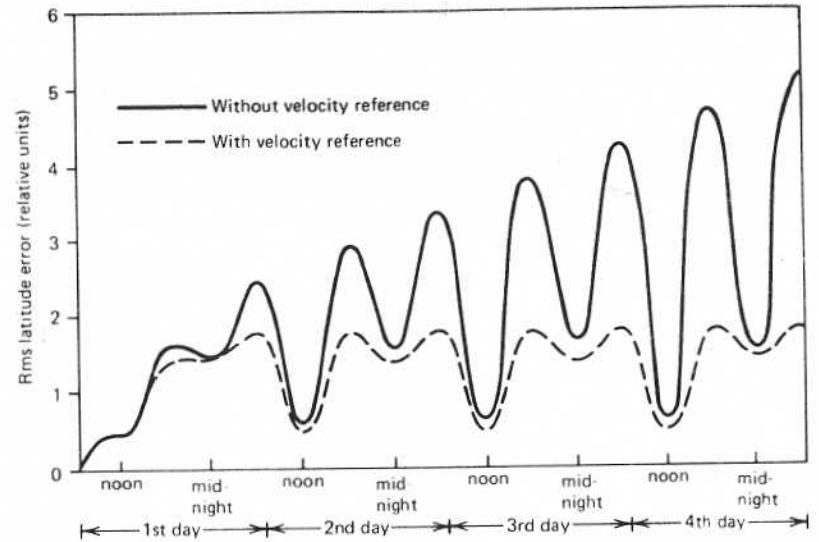


Figure 6.9. Rms errors with and without velocity reference information.

6.7 OFF-LINE SYSTEM ERROR ANALYSIS

In addition to its value as an estimator, the Kalman filter also provides a convenient means of system error analysis. In preliminary analysis and design, the analyst often needs to assess various competing designs relative to their effect on system accuracy. Recall that in the filter recursive equations, the error covariance \mathbf{P} is propagated along with the estimate $\hat{\mathbf{x}}$. As a matter of fact, this cannot be avoided because \mathbf{P} is needed for the gain computation and subsequent updating of the estimate. The reverse is not true though. The \mathbf{P} matrix can be propagated without forming the estimate. With a modest amount of algebra, one can write \mathbf{P}_{k+1} explicitly as a function of \mathbf{P}_k , and thus have a single recursive equation for the error covariance matrix (see Problem 6.6). This leads to no significant saving in the number of arithmetic steps needed in the loop, though, so the analyst usually programs the usual filter loop shown in Fig. 5.7 with the estimate computations omitted. The abbreviated loop usually used for analysis of the optimal estimator is summarized in Fig. 6.10. It should be emphasized here that the \mathbf{P} matrix does not depend on the actual measurement data received on any particular sample run. Thus, the Kalman filter in its pure form is not adaptive. The exception would be the case where the system has the capability of accepting or rejecting measurement data on-line based on some sort of validity check. Many systems are designed with this feature, and they might be said to be adaptive in this primitive sense.

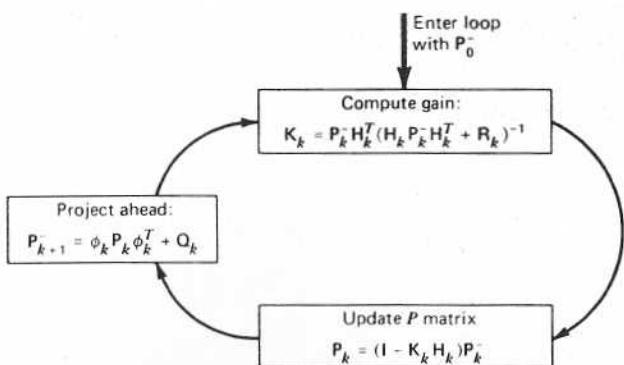


Figure 6.10. Recursive loop for propagating optimal error covariance matrix.

When the optimal gain is used in error analysis, the results represent the ideal situation where the filter model and the actual process model are identical. For purposes of comparison, it is also desirable to be able to assess the error in suboptimal systems. This can be done by cycling suboptimal gains through a truth-model loop that uses the appropriate \mathbf{P} -update equation. Recall that Eq. (5.4.11) is valid for any gain, optimal or suboptimal. The procedure for performing suboptimal error analysis is then summarized in Fig. 6.11. If you wish to make a comparison between optimal and suboptimal systems, the optimal loop of Fig. 6.10 must also be run in addition to the loops shown in Fig. 6.11. Usually the loops are all run in parallel, because some of the computations required are common to all loops. Be careful, though, to keep the gains and \mathbf{P} -matrix computations of each loop separated because they are all different (even though we have not attempted to distinguish among them in the notation in Figs. 6.10 and 6.11). It is a bit confusing, but there are three \mathbf{P} matrices to keep straight:

1. \mathbf{P} of the optimal estimator: This always yields the most optimistic results—that which might be expected if the filter model and the truth model match perfectly.
2. \mathbf{P} of the truth model with suboptimal gains: This usually yields more realistic results. It represents the actual errors to be expected when the filter model and the truth model are not the same.
3. \mathbf{P} of the suboptimal filter that generates the suboptimal gain sequence. This matrix has no significance in terms of the errors of anything. It can be viewed just as a necessary adjunct for computing the suboptimal gain sequence.

Note that the gain and \mathbf{P} matrices of the optimal and suboptimal models may not even have the same dimensions. Often the analyst wishes to assess the

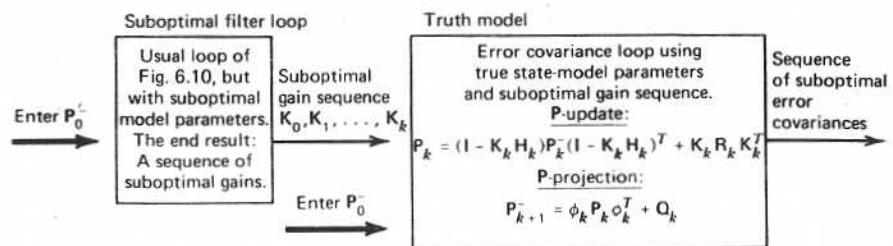


Figure 6.11. Block diagram for suboptimal error analysis.

degradation in performance caused by deletion of some of the less-important state variables in the state model. In this case, the suboptimal model is of lesser order than the truth model, and suboptimal gains for the deleted state variables are not generated by the suboptimal model. Thus, zero gains must be inserted appropriately in the truth-model calculations.

Note also that the \mathbf{P} -projection equation in the truth model with suboptimal gains is the same as that used in the optimal loop. This equation is valid for both loops because it depends only on the linear transformation of errors that occurs from step to step and the white properties of the w_k sequence. In both optimal and suboptimal situations, the estimation errors are propagated through the same identical truth model, and therefore the \mathbf{P} -projection equation is the same for both. A simple example that illustrates suboptimal error analysis follows.

Example 6.5 We return to Example 6.4 in which a random-walk process was incorrectly modeled as a random constant. One sample run of a random-walk process was used to demonstrate the divergence phenomenon. However, this can hardly be called proof of divergence. We could, of course, make many more runs using new sets of random numbers, and then average the results to find the rms error. This would be doing the analysis the hard way, though. With suboptimal variance analysis, we can obtain the rms error with one run! All we need to do is consider the random-walk model as the truth model and the random-constant model as the suboptimal system, and then cycle the suboptimal gains through the truth model as indicated in Fig. 6.11. This was done for the first 15 steps of this example, and the resulting rms error along with suboptimal gains is shown in Fig. 6.12. The optimal rms error is also shown for comparison. It can be seen that divergence does occur in this situation. In this example, it can be easily verified that the suboptimal filter approaches a steady-state condition where the error variance increases by a fixed amount with each step. Thus, the rms error increases as the square root of the number of steps as indicated in Fig. 6.12. ■

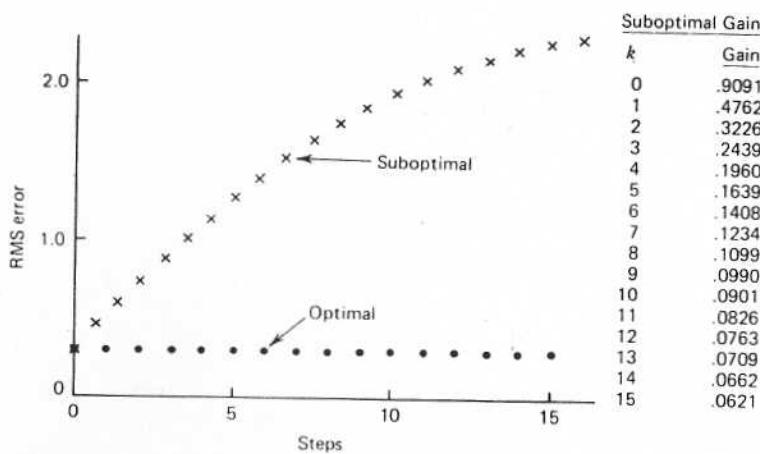


Figure 6.12. Suboptimal analysis for random-walk example.

6.8 RELATIONSHIP TO DETERMINISTIC LEAST SQUARES

Both Kalman and Wiener filtering are sometimes referred to as least-squares filtering (17,18,19). It was mentioned in Chapter 4 that this is somewhat of an oversimplification, because the criterion for optimization is minimum mean-square error and not the squared error in a deterministic sense. There is, however, a coincidental connection between Kalman/Wiener filtering and deterministic least squares, and this will now be demonstrated. The presentation here follows closely that of Sorenson (6).

Consider a set of m linear equations in \mathbf{x} specified in matrix form by

$$\mathbf{M}\mathbf{x} = \mathbf{b} \quad (6.8.1)$$

In Eq. (6.8.1) we think of \mathbf{M} and \mathbf{b} as being given, and \mathbf{x} is $(n \times 1)$, \mathbf{b} is $(m \times 1)$, and thus \mathbf{M} is $(m \times n)$. Let us assume that $m > n$, and that \mathbf{x} is overdetermined by the system of equations represented by Eq. (6.8.1). Thus no solution for \mathbf{x} will satisfy all equations. This situation arises frequently in physical experiments where redundant noisy measurements are made of linear combinations of fixed parameters. In such cases it is logical to ask, "What solution will best fit all the equations?" The term best must, of course, be defined and it is frequently defined to be the particular \mathbf{x} , say \mathbf{x}_{opt} , that minimizes the sum of the squared residuals. That is, move \mathbf{b} to the left side of Eq. (6.8.1) and substitute \mathbf{x}_{opt} for \mathbf{x} . This yields a residual vector $\boldsymbol{\epsilon}$ given by

$$\mathbf{M}\mathbf{x}_{\text{opt}} - \mathbf{b} = \boldsymbol{\epsilon} \quad (6.8.2)$$

and \mathbf{x}_{opt} is chosen such that $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$ is minimized. A perfect fit, of course, would make $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = 0$.

We can generalize at this point and consider a weighted sum of squared residuals specified by

$$[\text{Weighted sum of squared residuals}] = (\mathbf{M}\mathbf{x}_{\text{opt}} - \mathbf{b})^T \mathbf{W} (\mathbf{M}\mathbf{x}_{\text{opt}} - \mathbf{b}) \quad (6.8.3)$$

We assume that the weighting matrix \mathbf{W} is symmetric and positive definite and, hence, so is its inverse. If we wish equal weighting of the residuals, we simply let \mathbf{W} be the identity matrix. The problem now is to find the particular \mathbf{x} (i.e., \mathbf{x}_{opt}) that minimizes the weighted sum of the residuals. Toward this end, the expression given by Eq. (6.8.3) may be expanded and differentiated term-by-term and then set equal to zero.* This leads to

$$\begin{aligned} \frac{d}{d\mathbf{x}_{\text{opt}}} [\mathbf{x}_{\text{opt}}^T (\mathbf{M}^T \mathbf{W} \mathbf{M}) \mathbf{x}_{\text{opt}} - \mathbf{b}^T \mathbf{W} \mathbf{M} \mathbf{x}_{\text{opt}} - \mathbf{x}_{\text{opt}}^T \mathbf{M}^T \mathbf{W} \mathbf{b} + \mathbf{b}^T \mathbf{b}] \\ = 2(\mathbf{M}^T \mathbf{W} \mathbf{M}) \mathbf{x}_{\text{opt}} - (\mathbf{b}^T \mathbf{W} \mathbf{M})^T - \mathbf{M}^T \mathbf{W} \mathbf{b} = 0 \end{aligned} \quad (6.8.4)$$

Equation (6.8.4) may now be solved for \mathbf{x}_{opt} . The result is

$$\mathbf{x}_{\text{opt}} = [(\mathbf{M}^T \mathbf{W} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{W}] \mathbf{b} \quad (6.8.5)$$

and this is the solution of the deterministic least-squares problem.

* The derivative of a scalar s with respect to a vector \mathbf{x} is defined to be

$$\frac{ds}{d\mathbf{x}} = \begin{bmatrix} \frac{ds}{dx_1} \\ \frac{ds}{dx_2} \\ \vdots \\ \frac{ds}{dx_n} \end{bmatrix}$$

The two matrix differentiation formulas used to arrive at Eq. (6.8.4) are

$$\frac{d(\mathbf{x}^T \mathbf{A} \mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x} \quad (\text{for symmetric } \mathbf{A})$$

and

$$\frac{d(\mathbf{a}^T \mathbf{x})}{d\mathbf{x}} = \frac{d(\mathbf{x}^T \mathbf{a})}{d\mathbf{x}} = \mathbf{a}$$

Both of these formulas can be verified by writing out a few scalar terms of the matrix expressions and using ordinary differentiation methods.

Next consider the Kalman filter solution for the same measurement situation. The vector \mathbf{x} is assumed to be a constant, so the differential equation for \mathbf{x} is

$$\dot{\mathbf{x}} = 0 \quad (6.8.6)$$

The corresponding discrete model is then

$$\mathbf{x}_{k+1} = \mathbf{I} \cdot \mathbf{x}_k + \mathbf{0} \quad (6.8.7)$$

The measurement equation is

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \quad (6.8.8)$$

where \mathbf{z}_k and \mathbf{H}_k play the same roles as \mathbf{b} and \mathbf{M} in the deterministic problem. Since time is of no consequence, we assume all measurements occur simultaneously. Furthermore, we assume that we have no a priori knowledge of \mathbf{x} , so the initial $\hat{\mathbf{x}}_0^-$ will be zero and its associated error covariance will be ∞ . Therefore, using the alternative form of the Kalman filter (Section 6.2), we have

$$\begin{aligned} \mathbf{P}_0^{-1} &= (\infty)^{-1} + \mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0 \\ &= \mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0 \end{aligned} \quad (6.8.9)$$

The Kalman gain is then

$$\mathbf{K}_0 = (\mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0)^{-1} \mathbf{H}_0^T \mathbf{R}_0^{-1}$$

and the Kalman filter estimate of \mathbf{x} at $t = 0$ is

$$\hat{\mathbf{x}}_0 = [(\mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0)^{-1} \mathbf{H}_0^T \mathbf{R}_0^{-1}] \mathbf{z}_0 \quad (6.8.10)$$

This is the same identical expression obtained for \mathbf{x}_{opt} in the deterministic least-squares problem with \mathbf{R}_0^{-1} playing the role of the weighting matrix \mathbf{W} .

Let us now recapitulate the conditions under which the Kalman filter estimate coincides with the deterministic least-squares estimate. First, the system state vector was assumed to be a constant (the process dynamics are thus trivial). Second, we assumed the measurement sequence was such as to yield an overdetermined system of linear equations [otherwise $(\mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0)^{-1}$ will not exist]. And, finally, we assumed that we had no prior knowledge about the constant vector being estimated. This latter assumption is unusual because in many situations we have at least some a priori knowledge of the process being estimated. One of the things that distinguishes the Kalman filter from other estimators is the convenient way in which it accounts for this prior knowledge via the initial conditions of the recursive process. (This was used to good advantage in the power system relaying example of Section 6.5.) Of course, if there is truly no prior knowledge to use, the Kalman filter advantage is lost (in this respect), and it degenerates to a least-squares fit under the conditions just stated.

The coincidence in the deterministic least-squares and Kalman filter estimates is really rather remarkable. Remember, one solution was obtained by posing a *deterministic* optimization problem; the other by posing a similar *stochastic* problem. There is no reason *offhand* to think these two approaches would lead to identical solutions. Yet they do under certain circumstances. The circumstances may be generalized somewhat from those of this example, but not to the complete extent of the general process model used in the Kalman filter. [See Sorenson (18) for more on this point.] Thus, this happy coincidence in the two solutions will not always exist.

6.9 TELEPHONE LOAD FORECASTING EXAMPLE

It was mentioned previously that applications of Kalman filtering have been spreading into areas far afield from the original applications in the aerospace field. For example, the January 1982 issue of the *Bell System Technical Journal* was devoted entirely to the application of Kalman filtering to telephone load forecasting. C. R. Szelag (21) had a particularly interesting paper in this issue of the *BSTJ*, and we will examine briefly a simplified version of this application.

The load history for one of the telephone trunks considered by Szelag (21) is repeated here in Fig. 6.13. The "squares" in the figure indicate the actual load data experienced by the system during the 1975–79 period. (The circles may be ignored for the moment.) Clearly, the load exhibits a general yearly oscillatory trend superimposed on a linearly increasing trend. However, there is some noisiness in the trend because the oscillatory part is not exactly sinusoidal, nor is the linear part exactly linear, which indicates that this process might be modeled as the sum of oscillatory and linear components that are

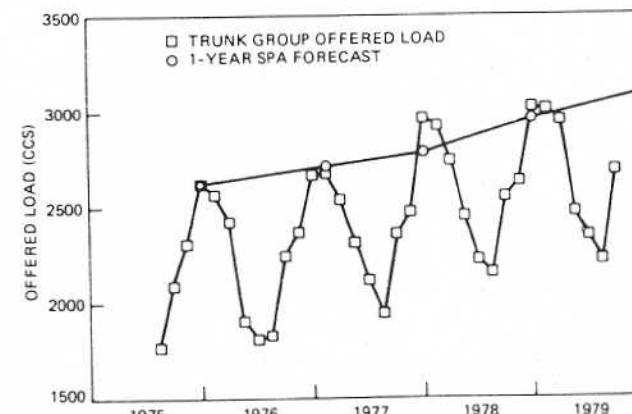


Figure 6.13. Telephone trunk load.

driven by white noise inputs. That is, we speculate that the following model might fit this situation reasonably well:

(a) Linear part:

$$\ddot{x} = f_1(t), \quad f_1(t) \text{ is white noise} \quad (6.9.1)$$

(b) Oscillatory part:

$$\ddot{y} + \omega^2 y = f_2(t), \quad f_2(t) \text{ white and independent of } f_1(t) \quad (6.9.2)$$

Next, Eqs. (6.9.1) and (6.9.2) may be written in vector form by defining state variables as follows:

$$x_1 = x \quad (6.9.3)$$

$$x_2 = \dot{x} \quad (6.9.4)$$

$$x_3 = y \quad (6.9.5)$$

$$x_4 = \dot{y} \quad (6.9.6)$$

The state equations are then

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -\omega^2 & 0 & 0 \end{array} \right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} f_1(t) \\ 0 \\ 0 \\ f_2(t) \end{bmatrix} \quad (6.9.7)$$

Let us now assume that the measurements are discrete and are uniformly spaced with a sampling interval of T . The discrete state model is then

$$\mathbf{x}_{k+1} = \left[\begin{array}{ccc|c} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos \omega T & \sin \omega T \\ 0 & 0 & -\sin \omega T & \cos \omega T \end{array} \right] \mathbf{x}_k + \mathbf{w}_k \quad (6.9.8)$$

The transition matrix ϕ_k is now determined as shown in component form in Eq. (6.9.8). The \mathbf{Q}_k matrix will depend on the amplitudes of the white-noise inputs $f_1(t)$ and $f_2(t)$. This parameter is usually the most elusive one of the model and often has to be determined by trial and error. Szelag (21) does not provide a numerical value for \mathbf{Q}_k in this case, but he does indicate that it was determined empirically.

The measurement equation is especially simple in this example because the total process being measured is just the sum of the linear and oscillatory parts. Thus z_k is scalar and we have

$$z_k = [1 \ 0 \ 1 \ 0] \mathbf{x}_k + v_k \quad (6.9.9)$$

Therefore \mathbf{H}_k is as shown in component form in Eq. (6.9.9). We would expect the R_k parameter associated with the measurement noise v_k to be relatively small in this case. As with the case of the \mathbf{Q}_k parameter, no numerical value for R_k was given in the referenced paper, so we have no check on this assumption. It is reasonable to assume, though, that the trunk load measurement is relatively accurate.

Let us now assume the ϕ_k , \mathbf{Q}_k , \mathbf{H}_k , and R_k parameters of the model have been determined. In this application the problem is to predict the load at the next step. That is, we have the past history of load measurements up to and including time t_k , and we wish to estimate the load at t_{k+1} . Thus it is the a priori estimate $\hat{\mathbf{x}}_{k+1}$ that is of primary interest in this application. (To be more specific, it is the sum of the first and third elements of the vector that is of interest.) The a priori or one-step prediction estimate is, of course, one of the normal computations in the usual Kalman filter algorithm; therefore, the recursive equations shown in Fig. 5.7 are directly applicable without any modification. There is a start-up problem, though, when using the algorithm either in real time or in an empirical evaluation of the estimator's merits using real-life data from past experience. With respect to the latter, the author chose to use the first 16 data points to initialize the filter. This then established an a posteriori estimate and associated error covariance matrix at step 16. These were then projected ahead to step 17, and the regular filter loop was continued to the end of the data stream.

Figure 6.14 shows the results of a sample run for the load history shown previously in Fig. 6.13. As before, the squares indicate the actual load data. The triangles are the a priori estimates of interest, and note that they do not begin until step 17. (The circles relate to the initialization and may be ignored for our purposes.) It can be seen that the Kalman filter does a reasonably good job of one-step prediction on this run.

A few further comments about initialization and other matters are in order before leaving this example. First, Szelag (21) chose to use a batch-processing weighted-least-squares approach in initializing his Kalman filter. In view of our comments about deterministic least squares in Section 6-8, it would appear that this could just as well have been done recursively. To do this, we would simply assume an artificially large \mathbf{P}^- matrix at the beginning of the data stream, and then the Kalman filter would automatically give the initial a priori estimate (e.g., zero) negligible weight. In this example the state vector being estimated is fourth order and the measurements are scalar; thus it would take the first four steps to bring the initial \mathbf{P} matrix down to normal size, so to speak. From that point on the filter would behave more or less normally, and the \mathbf{P} matrix should reach a near stationary condition by step 16. Assessment of the filter's accuracy as a one-step predictor could then begin with step 17.

We might also note that the model chosen here for tutorial purposes is relatively simple. Szelag notes in his paper that higher-order harmonics may be added to the model. Of course, this refinement is not without a price because

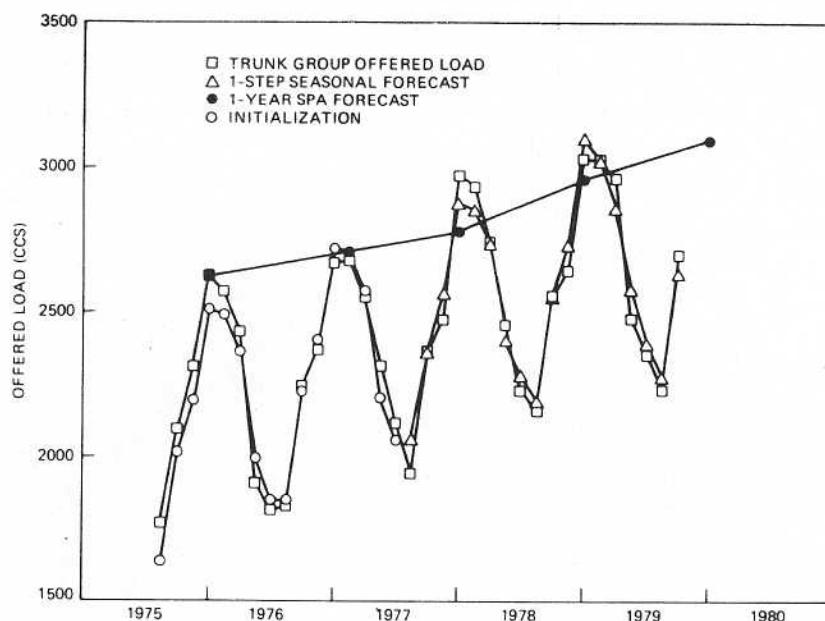


Figure 6.14. Telephone trunk load data along with one-step predictions.

each harmonic added to the model increases the dimensionality of the state vector by two units. Also, the input noise model could be refined by allowing white noise to enter the \dot{x}_2 and \dot{x}_4 equations directly, as well as in the \dot{x}_1 and \dot{x}_3 equations; see Eq. (6.9.7). This does not appear to be worth the effort in this example, though, in view of the fact that Q_k is being determined empirically. In any event, this is a new and novel application, and it is interesting to see Kalman filtering spread to fields outside of aerospace engineering.

6.10 DETERMINISTIC INPUTS

In many situations the random processes under consideration are driven by deterministic as well as random inputs. That is, the process equation may be of the form

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{Gw} + \mathbf{Bu} \quad (6.10.1)$$

where \mathbf{Bu} is the additional deterministic input. Since the system is linear we can use superposition and consider the random and deterministic responses separately. Thus the discrete Kalman filter equations are only modified slightly. The only change required is in the estimate projection equation. In this equation the

contribution due to \mathbf{Bu} must be properly accounted for. Using the same zero-mean argument as before, relative to the random response, we then have

$$\hat{\mathbf{x}}_{k+1}^- = \boldsymbol{\phi}_k \hat{\mathbf{x}}_k + 0 + \int_{t_k}^{t_{k+1}} \boldsymbol{\phi}(t_{k+1}, \tau) \mathbf{B}(\tau) \mathbf{u}(\tau) d\tau \quad (6.10.2)$$

where the integral term is the contribution due to \mathbf{Bu} in the interval (t_k, t_{k+1}) . The associated equation for \mathbf{P}_{k+1}^- is $(\boldsymbol{\phi}_k \mathbf{P}_k \boldsymbol{\phi}_k^T + \mathbf{Q}_k)$, as before, because the uncertainty in the deterministic term is zero. Also, the estimate update and associated covariance expressions (see Fig. 5.7) are unchanged, provided the deterministic contribution has been properly accounted for in computing the a priori estimate $\hat{\mathbf{x}}_k$.

Another way of accounting for the deterministic input is to treat the problem as a superposition of two entirely separate estimation problems, one deterministic and the other random. The deterministic one is trivial, of course, and the random one is not trivial. This complete separation approach is not necessary, though, provided one properly accounts for the deterministic contribution in the projection step.

Problems

6.1. The ψ equations that describe the error propagation in a slow-moving inertial navigation system were given in Section 6.1 [Eqs. (6.13) to (6.15)]. Let ψ_x, ψ_y, ψ_z be state vector elements x_1, x_2, x_3 and $\varepsilon_x, \varepsilon_y, \varepsilon_z$ be the system inputs. First, write the equations in the standard state-space form:

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{Bu}$$

Next, find the eigenvalues of the \mathbf{F} matrix. (Note: These describe the natural modes of the system.)

6.2. Consider a simplified version of Example 6.1 in which the gyro drifts are modeled simply as random constants. The system state vector then reduces to a 6-tuple. Also, assume the sampling interval for the measurements Δt is constant and small relative to the natural period of the system (i.e., 1 day, see Problem 6.1). Find the $\boldsymbol{\phi}_k$, \mathbf{Q}_k , \mathbf{H}_k , \mathbf{R}_k and initial condition parameters \mathbf{P}_0^- , $\hat{\mathbf{x}}_0^-$ for this system. Assume the earth-rate components Ω_x and Ω_z are known constants. Also, assume the initial uncertainties in ψ_x, ψ_y, ψ_z are all the same with variance σ_ψ^2 , and the corresponding initial variances for $\varepsilon_x, \varepsilon_y, \varepsilon_z$ are all σ_ε^2 . All initial cross-correlations are zero, and the measurement-error variances are σ_x^2 and σ_y^2 .

6.3. Is the system of Problem 6.2 observable? Recall from linear system theory that the test for observability is to form the matrix

$$\mathbf{M} = [\mathbf{H}^T, \mathbf{F}^T \mathbf{H}^T, (\mathbf{F}^T)^2 \mathbf{H}^T, \dots, (\mathbf{F}^T)^{n-1} \mathbf{H}^T]$$

and then test the rank of \mathbf{M} . If it is of full rank, the system is observable; if not, the system is not observable. (Answer: The system is not observable. This can also be seen by writing out the explicit solution of the state equations in terms of the initial condi-

tions. Certain linear combinations of state variables always appear together and are thus inseparable.)

6.4. Consider the measurement to be a 2-tuple $[z_1, z_2]^T$, and assume the measurement errors are correlated such that the \mathbf{R} matrix is of the form

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{bmatrix}$$

(a) Form a new measurement pair, z'_1 and z'_2 , as a linear combination of the original pair such that the errors in the new pair are uncorrelated. (*Hint:* First, let $z'_1 = z_1$ and then assume $z'_2 = c_1 z_1 + c_2 z_2$ and choose the constants c_1 and c_2 such that the new measurement errors are uncorrelated.)

(b) Find the \mathbf{R} matrix associated with the new \mathbf{z}' measurement vector.

6.5. The current model of the relaying example presented in Section 6.5 contained a nonstationary state variable with exponential properties. Assume the initial value of this state variable is a random variable and has a variance that is 100 times larger than the steady-state variance of the process. Sketch three typical sample functions of the process.

6.6. (a) Derive an explicit expression for \mathbf{P}_{k+1} in terms of \mathbf{P}_k and the ϕ , \mathbf{H} , \mathbf{Q} , \mathbf{R} , parameters. (The a priori error covariance matrix \mathbf{P}^- and gain \mathbf{K} should not appear in your final expression.)

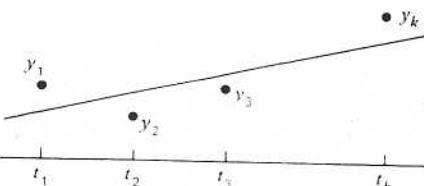
(b) Find a similar difference equation for \mathbf{P}_k^- in terms of \mathbf{P}_k^- and the system parameters.

6.7. Suboptimal filter analysis is often used in sensitivity analysis. This is where the analyst wishes to assess the sensitivity of the system error to changes in certain parameters. In the random-walk example of Section 6.7, the true value of Q was said to be 1, and the rms error was computed accordingly. Assume a 10 percent variation of the true Q on either side of the nominal value of 1, and find the corresponding changes in rms estimation error. (*Note:* Convergence to the steady-state condition is quite rapid in this case. Thus, the analysis can be carried out easily with a hand-held calculator.)

6.8. Consider the sequence of pairs of real numbers $(y_1, t_1), (y_2, t_2), \dots, (y_k, t_k)$ where t_1, t_2, \dots, t_k are arranged in ascending order. Now think of y_1, y_2, \dots, y_k as noisy samples of the exact linear relationship

$$y = at + b$$

as shown in the accompanying figure. Consider the problem of estimating a and b , based on the sequence of samples y_1, y_2, \dots, y_k .



Problem 6.8

(a) Find the Kalman-filter estimate of a and b . Assume there is no prior knowledge of a and b , and that all measurement errors are uncorrelated and have equal variances σ^2 .

(b) Demonstrate for the case of three samples that the Kalman estimate is the same as would be obtained if one were to do a deterministic least-squares fit of a linear function to the three data points y_1, y_2, y_3 .

6.9. Consider an elementary physics experiment that is intended to measure the gravity constant g . A mass is released at $t = 0$ in a vertical, evacuated column, and multiple-exposure photographs of the falling mass are taken at .05-sec intervals beginning at $t = .05$ sec. A sequence of N such exposures is taken, and then the position of the mass at each time is read from a scale in the photograph. There will be experimental errors for a number of reasons; let us assume they are random (i.e., not systematic) and are such that the statistical uncertainties in all position readings are all the same and that the standard deviation of these is σ .

(a) Consider g to be an unknown constant with no prior knowledge as to its value. Develop a Kalman filter model for processing the position measurements, treating g as the state of the system. The objective is, of course, to estimate g .

(b) Suppose we choose a 3-state dynamical model with position, velocity, and g as the state variables. After assimilating one measurement, show that the estimate of g is the same as that obtained in part (a). Also show that the Kalman filter estimates of position and velocity are the same as those obtained from the usual physics formulas

$$\text{position} = \frac{1}{2}(\text{acc.}) \times (\text{time})^2$$

$$\text{velocity} = (\text{acc.}) \times (\text{time})$$

where \hat{g} is used for acceleration.

REFERENCES CITED IN CHAPTER 6

- B. E. Bona, and R. J. Smay, "Optimum Reset of Ship's Inertial Navigation System," *IEEE Trans. on Aerospace and Electronic Systems*, AES-2: No. 4, 409-414 (July 1966).
- G. R. Pitman (Ed.), *Inertial Guidance*, New York: Wiley, 1962.
- J. C. Pinson, "Inertial Guidance for Cruise Vehicles," in *Guidance and Control of Aerospace Vehicles*, C. T. Leondes, (Ed.), New York: McGraw-Hill, 1963.
- J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*, New York: McGraw-Hill, 1969.
- A. Gelb (Ed.), *Applied Optimal Estimation*, Cambridge, Mass.: MIT Press, 1974.
- H. W. Sorenson, "Kalman Filtering Techniques," in *Advances in Control Systems* (Vol. 3), C. T. Leondes, (Ed.), New York: Academic Press, 1966.
- R. J. Milliken, and C. J. Zoller, "Principle of Operation of NAVSTAR and System Characteristics," *Navigation, Journal of the Inst. of Navigation*, 25: No. 2 (Special Issue on GPS), 95-106 (Summer 1978).
- E. M. Copps, G. J. Geier, W. C. Fidler, and P. A. Grundy, "Optimal Processing of GPS Signals," *Navigation, Journal of the Inst. of Navigation*, 27: No. 3, 171-182 (Fall 1980).
- R. G. Brown, "Integrated Navigation Systems and Kalman Filtering: A Perspective," *Navigation, Jour. of the Inst. of Navigation*, 19: No. 4, 355-362 (Winter 1972-73).

10. J. D. Salisbury, "Comments on Integrated Navigation Systems and Kalman Filtering: A Perspective," *Navigation, Jour. of the Inst. of Navigation*, 20: No. 2, 190 (Summer 1973).
11. P. S. Maybeck, *Stochastic Models, Estimation and Control*, Vol. 1, New York: Academic Press, 1979.
12. R. G. Brown, and D. T. Friest, "Optimization of a Hybrid Inertial Solar-Tracker Navigation System," *1964 IEEE International Convention Record*, Part 7: 121-135.
13. A. A. Girgis, "Application of Kalman Filtering in Computer Relaying of Power Systems," Ph.D. dissertation, Iowa State University, Ames, 1981.
14. A. A. Girgis, and R. G. Brown, "Application of Kalman Filtering in Computer Relaying," *IEEE Trans. on Power Apparatus and Systems, PAS-100*: No. 7, 3387-3397 (July 1981).
15. G. R. Marner, "Automatic Radio-Celestial Navigation," *Journal of the Institute of Navigation (British) XII*: Nos. 3 and 4, 249-259 (July/Oct. 1959).
16. N. Bowditch, *American Practical Navigator*, H. O. Pub. No. 9, Washington, D.C.: U.S. Government Printing Office, 1966.
17. H. W. Bode, and C. E. Shannon, "A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory," *Proc. I.R.E.*, 38, 417-424 (April 1950).
18. H. W. Sorenson, "Least-Squares Estimation: From Gauss to Kalman," *IEEE Spectrum*, 7, 63-68 (July 1970).
19. T. Kailath, "A View of Three Decades of Linear Filtering Theory," *IEEE Trans. on Information Theory, IT-20*: No. 2, 146-181 (March 1974).
20. C. T. Leondes, (Ed.), *Theory and Applications of Kalman Filtering*, North Atlantic Treaty Organization AGARD Report No. 139 (February 1970).
21. C. R. Szelag, "A Short-Term Forecasting Algorithm for Trunk Demand Servicing," *The Bell System Technical Journal*, 61: No. 1, 67-96 (January 1982).

CHAPTER 7

The Continuous Kalman Filter

About a year after his paper on discrete-data filtering, R. E. Kalman coauthored a second paper with R. S. Bucy on continuous filtering (1). This paper also proved to be a milestone in the area of optimal filtering. Our approach here will be somewhat different from theirs, in that we will derive the continuous filter equations as a limiting case of the discrete equations as the step size becomes small. Philosophically, it is of interest to note that we begin with the discrete equations and then go to the continuous equations. So often in numerical procedures, we begin with the continuous dynamical equations; these are then discretized and the discrete equations become approximations of the continuous dynamics. Not so with the Kalman filter! The discrete equations are exact and stand in their own right, provided, of course, that the difference equation model of the process is exact and not an approximation.

The continuous Kalman filter is probably not as important in applications as the discrete filter, especially in real-time systems. However, the continuous filter is important for both conceptual and theoretical reasons, so we now proceed with its derivation.

7.1 TRANSITION FROM THE DISCRETE TO CONTINUOUS FILTER EQUATIONS

First, we assume the process and measurement models to be of the form:

$$\text{Process model: } \dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{w} \quad (7.1.1)$$

$$\text{Measurement model: } \mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v} \quad (7.1.2)$$

where

$$E[\mathbf{w}(t)\mathbf{w}^T(\tau)] = \mathbf{Q}\delta(t - \tau) \quad (7.1.3)$$

$$E[\mathbf{v}(t)\mathbf{v}^T(\tau)] = \mathbf{R}\delta(t - \tau) \quad (7.1.4)$$

$$E[\mathbf{w}(t)\mathbf{v}^T(\tau)] = 0 \quad (7.1.5)$$

We note that in Eqs. (7.1.1) and (7.1.2), \mathbf{F} , \mathbf{G} and \mathbf{H} may be time varying. Also, Eqs. (7.1.3) to (7.1.5) simply say that $\mathbf{w}(t)$ and $\mathbf{v}(t)$ are uncorrelated vector

white noise processes. \mathbf{Q} and \mathbf{R} play roles similar to \mathbf{Q}_k and \mathbf{R}_k in the discrete filter, but they do not have the same numerical values. The relationships between the corresponding discrete and continuous filter parameters will be derived presently.

Recall that for the discrete filter,

$$\mathbf{Q}_k = E[\mathbf{w}_k \mathbf{w}_k^T] \quad (7.1.6)$$

$$\mathbf{R}_k = E[\mathbf{v}_k \mathbf{v}_k^T] \quad (7.1.7)$$

To make the transition from the discrete to continuous case, we need the relations between \mathbf{Q}_k and \mathbf{R}_k and the corresponding \mathbf{Q} and \mathbf{R} for a small step size Δt . Looking at \mathbf{Q}_k first, and referring to Eq. (5.3.6), we note that $\phi \approx \mathbf{I}$ for small Δt and thus

$$\mathbf{Q}_k \approx \iint_{\text{small } \Delta t} \mathbf{G}(u) E[\mathbf{w}(u) \mathbf{w}^T(v)] \mathbf{G}^T(v) du dv \quad (7.1.8)$$

Next, substituting Eq. (7.1.3) into (7.1.8) and integrating over the small Δt interval yields

$$\mathbf{Q}_k = \mathbf{G} \mathbf{Q} \mathbf{G}^T \Delta t \quad (7.1.9)$$

The derivation of the equation relating \mathbf{R}_k and \mathbf{R} is more subtle. In the continuous model $\mathbf{v}(t)$ is white, so simple sampling of $\mathbf{z}(t)$ leads to measurement noise with infinite variance. Hence, in the sampling process, we have to imagine averaging the continuous measurement over the Δt interval to get an equivalent discrete sample. This is justified because \mathbf{x} is not white and may be approximated as a constant over the interval. Thus we have

$$\begin{aligned} \mathbf{z}_k &= \frac{1}{\Delta t} \int_{t_{k-1}}^{t_k} \mathbf{z}(t) dt = \frac{1}{\Delta t} \int_{t_{k-1}}^{t_k} [\mathbf{H}\mathbf{x}(t) + \mathbf{v}(t)] dt \\ &\approx \mathbf{H}\mathbf{x}_k + \frac{1}{\Delta t} \int_{t_{k-1}}^{t_k} \mathbf{v}(t) dt \end{aligned} \quad (7.1.10)$$

The discrete-to-continuous equivalence is then

$$\mathbf{v}_k = \frac{1}{\Delta t} \int_{\text{small } \Delta t} \mathbf{v}(t) dt \quad (7.1.11)$$

From Eq. (7.1.7) we have

$$E[\mathbf{v}_k \mathbf{v}_k^T] = \mathbf{R}_k = \frac{1}{\Delta t^2} \iint_{\text{small } \Delta t} E[\mathbf{v}(u) \mathbf{v}^T(v)] du dv \quad (7.1.12)$$

Substituting Eq. (7.1.4) into (7.1.12) and integrating yields the desired relationship

$$\mathbf{R}_k = \frac{\mathbf{R}}{\Delta t} \quad (7.1.13)$$

7.1 TRANSITION FROM THE DISCRETE TO CONTINUOUS FILTER EQUATIONS 255

At first glance, it may seem strange to have the discrete measurement error approach ∞ as $\Delta t \rightarrow 0$. However, this is offset by the sampling rate becoming infinite at the same time.

In making the transition from the discrete to continuous case, we first note from the error covariance projection equation (i.e., $\mathbf{P}_{k+1}^- = \phi_k \mathbf{P}_k \phi_k^T + \mathbf{Q}_k$) that $\mathbf{P}_{k+1}^- \rightarrow \mathbf{P}_k$ as $\Delta t \rightarrow 0$. Thus, we do not need to distinguish between a priori and a posteriori \mathbf{P} matrices in the continuous filter. We proceed with the derivation of the continuous gain expression. Recall that the discrete Kalman gain is given by (see Fig. 5.7)

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k + \mathbf{R}_k)^{-1} \quad (7.1.14)$$

Using Eq. (7.1.13) and noting that $\mathbf{R}/\Delta t \gg \mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k^T$ leads to

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k + \mathbf{R}/\Delta t)^{-1} \approx \mathbf{P}_k^- \mathbf{H}_k^T \mathbf{R}^{-1} \Delta t$$

We can now drop the subscripts and the super minus on the right side and we obtain

$$\mathbf{K}_k = (\mathbf{P} \mathbf{H}^T \mathbf{R}^{-1}) \Delta t \quad (7.1.15)$$

We define the continuous Kalman gain as the coefficient of Δt in Eq. (7.1.15), that is,

$$\mathbf{K} \stackrel{\Delta}{=} \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} \quad (7.1.16)$$

Next, we look at the error covariance equation. From the projection and update equations (Fig. 5.7), we have

$$\begin{aligned} \mathbf{P}_{k+1}^- &= \phi_k \mathbf{P}_k \phi_k^T + \mathbf{Q}_k \\ &= \phi_k (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^- \phi_k^T + \mathbf{Q}_k \\ &= \phi_k \mathbf{P}_k^- \phi_k^T - \phi_k \mathbf{K}_k \mathbf{H}_k \mathbf{P}_k^- \phi_k^T + \mathbf{Q}_k \end{aligned} \quad (7.1.17)$$

We now approximate ϕ_k as $\mathbf{I} + \mathbf{F}\Delta t$ and note from Eq. (7.1.15) that \mathbf{K}_k is of the order of Δt . After neglecting higher-order terms in Δt , Eq. (7.1.17) becomes

$$\mathbf{P}_{k+1}^- = \mathbf{P}_k^- + \mathbf{F} \mathbf{P}_k^- \Delta t + \mathbf{P}_k^- \mathbf{F}^T \Delta t - \mathbf{K}_k \mathbf{H}_k \mathbf{P}_k^- + \mathbf{Q}_k \quad (7.1.18)$$

We next substitute the expressions for \mathbf{K}_k and \mathbf{Q}_k , Eqs. (7.1.15) and (7.1.9), and form the finite difference expression

$$\frac{\mathbf{P}_{k+1}^- - \mathbf{P}_k^-}{\Delta t} = \mathbf{F} \mathbf{P}_k^- + \mathbf{P}_k^- \mathbf{F}^T - \mathbf{P}_k^- \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}_k \mathbf{P}_k^- + \mathbf{G} \mathbf{Q} \mathbf{G}^T \quad (7.1.19)$$

Finally, passing to the limit as $\Delta t \rightarrow 0$ and dropping the subscripts and super-minus leads to the matrix differential equation

$$\dot{\mathbf{P}} = \mathbf{F} \mathbf{P} + \mathbf{P} \mathbf{F}^T - \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{P} + \mathbf{G} \mathbf{Q} \mathbf{G}^T \quad (7.1.20)$$

$$\mathbf{P}(0) = \mathbf{P}_0$$

Next, consider the state estimation equation. Recall the discrete equation is

$$\hat{x}_k = \hat{x}_{k-1} + K_k(z_k - H_k \hat{x}_{k-1}) \quad (7.1.21)$$

We now note that $\hat{x}_{k-1} = \phi_{k-1} \hat{x}_{k-1}$. Thus, Eq. (7.1.21) can be written as

$$\hat{x}_k = \phi_{k-1} \hat{x}_{k-1} + K_k(z_k - H_k \phi_{k-1} \hat{x}_{k-1}) \quad (7.1.22)$$

Again, we approximate ϕ as $I + F\Delta t$. Then, neglecting higher-order terms in Δt and noting that $K_k = K\Delta t$ leads to

$$\dot{\hat{x}}_k - \hat{x}_{k-1} = F\hat{x}_{k-1} \Delta t + K\Delta t(z_k - H_k \hat{x}_{k-1}) \quad (7.1.23)$$

Finally, dividing by Δt , passing to the limit, and dropping the subscripts yields the differential equation

$$\dot{\hat{x}} = F\hat{x} + K(z - H\hat{x}) \quad (7.1.24)$$

Equations (7.1.16), (7.1.20), and (7.1.24) comprise the continuous Kalman filter equations and these are summarized in Fig. 7.1. If the filter were to be implemented on-line, note that certain equations would have to be solved in real time as indicated in Fig. 7.1. Theoretically, the differential equation for P could be solved off-line, and the gain profile could be stored for later use on-line. However, the main \hat{x} equation must be solved on-line, because $z(t)$, that is, the noisy measurement, is the input to the differential equation.

The continuous filter equations as summarized in Fig. 7.1 are innocent looking because they are written in matrix form. They should be treated with

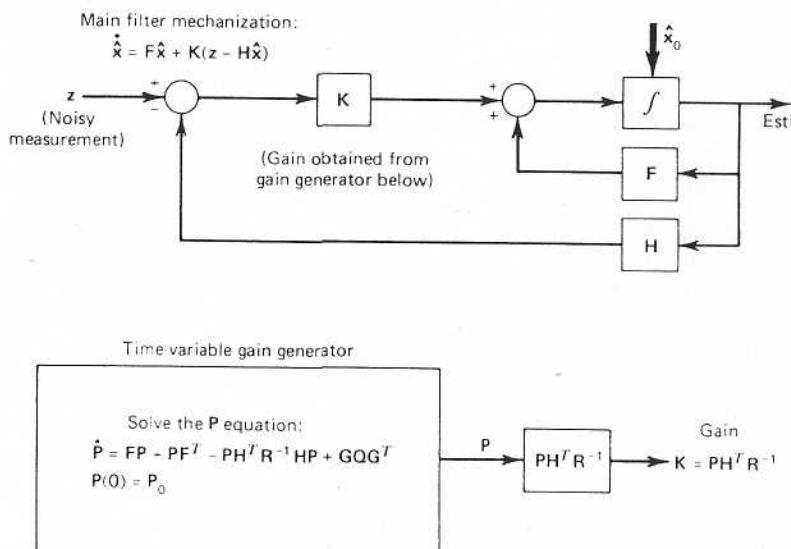


Figure 7.1. On-line block diagram for the continuous Kalman filter.

respect, though. It does not take much imagination to see the degree of complexity that results when they are written out in scalar form. If the dimensionality is high, an analog implementation is completely unwieldy.

Note that the error covariance equation must be solved in order to find the gain, just as in the discrete case. In the continuous case, though, a differential rather than difference equation must be solved. Furthermore, the differential equation is nonlinear because of the $PH^T R^{-1} HP$ term, which complicates matters. This will be explored further in the next section.

7.2 SOLUTION OF THE MATRIX RICCATI EQUATION

The error covariance equation

$$\dot{P} = FP + PF^T - PH^T R^{-1} HP + GQG^T \quad (7.2.1)$$

$$P(0) = P_0$$

is a special form of nonlinear differential equation known as the matrix Riccati equation. This equation has been studied extensively, and an analytical solution exists for the constant-parameter case. The general procedure is to transform the single nonlinear equation into a system of two simultaneous linear equations; and, of course, analytical solutions exist for linear differential equations with constant coefficients. Toward this end we assume P can be written in product form as

$$P = XZ^{-1}, \quad Z(0) = I \quad (7.2.2)$$

or

$$PZ = X \quad (7.2.3)$$

Differentiating both sides of Eq. (7.2.3) leads to

$$\dot{P}Z + P\dot{Z} = \dot{X} \quad (7.2.4)$$

Next, we substitute \dot{P} from Eq. (7.2.1) into Eq. (7.2.4) and obtain

$$(PF + PF^T - PH^T R^{-1} HP + GQG^T)Z + P\dot{Z} = \dot{X} \quad (7.2.5)$$

Rearranging terms and noting that $PZ = X$ leads to

$$P(F^T Z - H^T R^{-1} H X + \dot{Z}) + (FX + GQG^T Z - \dot{X}) = 0 \quad (7.2.6)$$

Note that if both terms in parentheses in Eq. (7.2.6) are set equal to zero, equality is satisfied. Thus, we have the pair of linear differential equations

$$\dot{X} = FX + GQG^T Z \quad (7.2.7)$$

$$\dot{Z} = H^T R^{-1} H X - F^T Z \quad (7.2.8)$$

with initial conditions

$$\begin{aligned} \mathbf{X}(0) &= \mathbf{P}_0 \\ \mathbf{Z}(0) &= \mathbf{I} \end{aligned} \quad (7.2.9)$$

These can now be solved by a variety of methods, including Laplace transforms. Once \mathbf{P} is found, the gain \mathbf{K} is obtained as $\mathbf{P}\mathbf{H}^T\mathbf{R}^{-1}$, and the filter parameters are determined. An example illustrates the procedure.

Example 7.1 We now return to the continuous filter problem considered previously in Example 4.5. The problem was solved there using Wiener methods, and we now wish to apply Kalman filtering methods. Let the signal and noise be statistically independent with autocorrelation functions

$$R_s(\tau) = e^{-|\tau|}, \quad [\text{or } S_s(s) = \frac{2}{-s^2 + 1}] \quad (7.2.10)$$

$$R_n(\tau) = \delta(\tau), \quad (\text{or } S_n = 1) \quad (7.2.11)$$

Since this is a one-state system, x is a scalar. Let x equal the signal. The additive measurement noise is white and thus no augmentation of the state vector is required. The process and measurement models are then

$$\dot{x} = -x + \sqrt{2}w, \quad w = \text{unity white noise} \quad (7.2.12)$$

$$z = x + v, \quad v = \text{unity white noise} \quad (7.2.13)$$

Thus, the system parameters are

$$F = -1, \quad G = \sqrt{2}, \quad Q = 1, \quad R = 1, \quad H = 1$$

The differential equations for X and Z are then

$$\begin{aligned} \dot{X} &= -X + 2Z, \quad X(0) = P_0 \\ \dot{Z} &= X + Z, \quad Z(0) = 1 \end{aligned} \quad (7.2.14)$$

Equations (7.2.14) may be solved readily using Laplace-transform techniques. The result is

$$\begin{aligned} X(t) &= P_0 \cosh \sqrt{3}t + \frac{(2 - P_0)}{\sqrt{3}} \sinh \sqrt{3}t \\ Z(t) &= \cosh \sqrt{3}t + \frac{(P_0 + 1)}{\sqrt{3}} \sinh \sqrt{3}t \end{aligned} \quad (7.2.15)$$

The solution for P may now be formed as $P = XZ^{-1}$:

$$P = \frac{P_0 \cosh \sqrt{3}t + \frac{(2 - P_0)}{\sqrt{3}} \sinh \sqrt{3}t}{\cosh \sqrt{3}t + \frac{(P_0 + 1)}{\sqrt{3}} \sinh \sqrt{3}t} \quad (7.2.16)$$

Once P is found, the gain \mathbf{K} is given by

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T\mathbf{R}^{-1}$$

and the filter yielding \hat{x} is determined.

Obviously there should be a correspondence between this solution and the one obtained by Wiener methods. The general connection between the two methods will be discussed in more detail in Section 7.6; here we simply consider a limiting case check as $t \rightarrow \infty$. This should yield the same steady-state (stationary) solution obtained previously with Wiener methods. Letting $t \rightarrow \infty$ in Eq. (7.2.16) and noting that $P_0 = 1$ yields

$$P \rightarrow \frac{1 \cdot e^{\sqrt{3}t} + \frac{2 - 1}{\sqrt{3}} e^{\sqrt{3}t}}{e^{\sqrt{3}t} + \frac{1 + 1}{\sqrt{3}} e^{\sqrt{3}t}} = \sqrt{3} - 1 \quad (7.2.17)$$

The Kalman filter block diagram for this example is then as shown in Fig. 7.2. This can be systematically reduced to yield the following overall transfer function relating \hat{x} to z :

$$G(s) = \frac{\text{Laplace transform of } \hat{x}}{\text{Laplace transform of } z} = \frac{\sqrt{3} - 1}{s + \sqrt{3}} \quad (7.2.18)$$

This is the same result obtained using Wiener methods. ■

7.3 CORRELATED MEASUREMENT AND PROCESS NOISE

Thus far we have considered w and v to be uncorrelated. However, occasionally we encounter applications where they are correlated. We will now see how the recursive equations can be modified to accommodate this situation. The

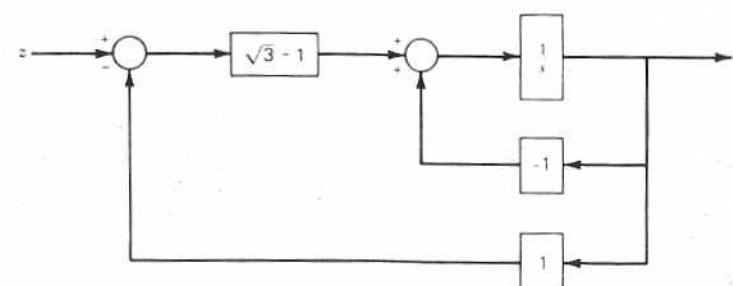


Figure 7.2. Stationary Kalman filter for Example 7.1.

process and measurement models are (as before)

$$\dot{x} = Fx + Gw \quad (7.3.1)$$

$$z = Hx + v \quad (7.3.2)$$

where

$$\begin{aligned} E[w(t)w^T(\tau)] &= Q\delta(t-\tau) \\ E[v(t)v^T(\tau)] &= R\delta(t-\tau) \end{aligned} \quad \text{(as before)} \quad (7.3.3)$$

and

$$E[w(t)v^T(\tau)] = C\delta(t-\tau) \quad (\text{rather than zero, as before}) \quad (7.3.4)$$

Our general approach is to form a new process model whose input noise is uncorrelated with v . We first note that $z - Hx - v$ is zero, and thus $D(z - Hx - v)$ may be added to the right side of Eq. (7.3.1). Thus we have

$$\dot{x} = Fx + Gw + D(z - Hx - v) \quad (7.3.5)$$

where the constant D will be chosen presently. However, first we rearrange the terms of Eq. (7.3.5) to obtain a new process model.

$$\text{New process model: } \dot{x} = (F - DH)x + Dz + (Gw - Dv) \quad (7.3.6)$$

The random process x is the same as before, but Eq. (7.3.6) shows that we can think of x as a superposition of two responses. One part is due to $Dz(t)$, which we will treat as if it were a known explicit input. (It is observed and available directly as a measurement.) The other part of the response is due to $(Gw - Dv)$, and this is a white noise input that is not observed (directly, at least). For lack of better names, we will refer to these two component responses as the explicit and random parts, and they will be denoted as x_e and x_r . The total response is then

$$x = x_e + x_r \quad (7.3.7)$$

Now, in the new process model, we wish to choose D such that $(Gw - Dv)$ and v are uncorrelated, that is,

$$E[(Gw(t) - Dv(t))v(\tau)^T] = 0 \quad (7.3.8)$$

or

$$E[Gw(t)v(\tau)^T] = E[Dv(t)v(\tau)^T] \quad (7.3.9)$$

Next, substituting Eqs. (7.3.3) and (7.3.4) into Eq. (7.3.9) leads to

$$GC\delta(t-\tau) = DR\delta(t-\tau) \quad (7.3.10)$$

We now see that if we choose D to be

$$D = GCR^{-1} \quad (7.3.11)$$

the desired decorrelation is effected. The new process model now satisfies all the necessary conditions imposed in Section 7.1, so the error covariance expression may be written as

$$\dot{P} = (F - DH)P + P(F - DH)^T - PH^T R^{-1} HP + Q' \quad (7.3.12)$$

where Q' is defined by

$$E\{[Gw(t) - Dv(t)][Gw(\tau) - DV(\tau)]^T\} = Q'\delta(t-\tau) \quad (7.3.13)$$

Expanding Eq. (7.3.13) and using Eqs. (7.3.3), (7.3.4), and (7.3.11) leads to

$$Q' = G(Q - CR^{-1}C^T)G^T \quad (7.3.14)$$

The equation for \dot{P} may now be rewritten as

$$\dot{P} = (F - DH)P + P(F - DH)^T - PH^T R^{-1} HP + G(Q - CR^{-1}C^T)G^T \quad (7.3.15)$$

The expression for gain in the new model is then

$$K' = PH^T R^{-1} \quad (7.3.16)$$

and the motive for using the prime will be made clear shortly.

We now look at the equation for the estimate. Just as in the case of the process itself, we wish to think of the estimate as a superposition of an explicit part, which is an estimate of x_e , and another part, which is an estimate of the random component x_r . Thus we have for the two estimates

$$\dot{\hat{x}}_e = (F - DH)\hat{x}_e + Dz(t) \quad (7.3.17)$$

$$\dot{\hat{x}}_r = (F - DH)\hat{x}_r + K'(z_r - H\hat{x}_r) \quad (7.3.18)$$

The error associated with \hat{x}_e is, of course, zero because $z(t)$ is known, that is, in real time it is the total measurement and it is known exactly. The measurement of x_r has been denoted as z_r , and this requires some explanation. From the basic measurement relationship, we have

$$z = Hx + v = Hx_e + Hx_r + v$$

or

$$(z - Hx_e) = Hx_r + v \quad (7.3.19)$$

Thus, $(z - Hx_e)$ must be considered as the noisy measurement of x_r . This is denoted as z_r in Eq. (7.3.18). Thus, making the substitution into Eq. (7.3.18) leads to

$$\dot{\hat{x}}_r = (F - DH)\hat{x}_r + K'(z - H\hat{x}_e - H\hat{x}_r) \quad (7.3.20)$$

Now, adding Eqs. (7.3.17) and (7.3.20), and noting that $\hat{x} = \hat{x}_e + \hat{x}_r$, yields

$$\dot{\hat{x}} = (F - DH)\hat{x} + (D + K')z - K'H\hat{x} = F\hat{x} + (D + K')(z - H\hat{x}) \quad (7.3.21)$$

It can be seen from the form of Eq. (7.3.21) that $(D + K')$ plays the role of "gain" in the estimation equation for the total \hat{x} quantity. The error associated with the \hat{x}_e component is zero, so the P equation, which was derived for the \hat{x} component, also applies to the total \hat{x} . From Eqs. (7.3.11) and (7.3.16), we see that the gain $D + K'$ is just $(PH^T + GC)R^{-1}$. Thus, the final estimation equations for \hat{x} may be summarized as

(1) Estimation equation:

$$\dot{\hat{x}} = F\hat{x} + K(z - H\hat{x}), \quad \hat{x}(0) = x_0 \quad (7.3.22)$$

(2) Gain equation:

$$K = (PH^T + GC)R^{-1} \quad (7.3.23)$$

(3) Error covariance equation:

$$\begin{aligned} \dot{P} &= (F - DH)P + P(F - DH)^T - PH^T R^{-1} HP \\ &\quad + G(Q - CR^{-1}C^T)G^T \end{aligned} \quad (7.3.24)$$

$$P(0) = P_0$$

where

$$D = GCR^{-1} \quad (7.3.25)$$

We note that if C is zero, the above estimation equations reduce to the previous equations for uncorrelated w and v . This is as expected. Also, note that in the process of deriving the filter equations for the correlated case, it was necessary to consider a superposition of explicit and random responses. Since this can always be done in a linear system, the addition of an explicit or deterministic driving function to the process equation presents no particular problem. We simply treat it separately and note that it contributes zero error to the estimate (see Problem 7.9). An example illustrating the use of the equations for correlated w and v will be presented in the next section.

7.4 COLORED MEASUREMENT NOISE

In the equations for the continuous Kalman filter, the inverse of the R matrix appears in both the gain and error covariance expressions. If the measurement noise is colored, rather than white, it must be modeled with additional state variables. This leaves zero for the white component, and thus R^{-1} will not exist. This leads to obvious difficulty. Of course, one remedy would be to add a small white component artificially, and then proceed on with the filter design using the usual equations. However, this is begging the issue. There are legitimate situations where one simply does not want to model any part of the measurement error as white noise.

A general solution to the colored measurement noise problem was first presented by Bryson and Johansen (2). Our approach here will be slightly

different and, we hope, more intuitively satisfying. If R^{-1} does not exist, we have a situation where certain linear combinations of the state variables are being measured perfectly. Obviously, if a particular state variable is known perfectly, it can be removed from the estimation problem. There is certainly no need for filtering something that is already free of corrupting noise. Therefore, our general approach will be to remove those quantities that are known perfectly from those that are not, and then solve the remaining estimation problem. This usually necessitates a linear transformation of the original state vector in order to decouple the known states from the others. This adds to the algebra, but it certainly presents no theoretical problem because the state estimate transforms with exactly the same transformation as the state itself.

Example 7.2 Consider the problem of separating an additive combination of two independent Markov processes with identical spectral functions. We know the solution calls for a trivial filter with a gain function of $\frac{1}{2}$. Let the autocorrelation functions of the signal and noise be

$$R_s(\tau) = e^{-|\tau|}, \quad R_n(\tau) = e^{-|\tau|} \quad (7.4.1)$$

The measurement z is the additive combination

$$z = s + n \quad (7.4.2)$$

Now the obvious way to model the continuous Kalman filter is to let s be x_1 and n be x_2 , and then we have

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad (7.4.3)$$

$$[z] = [1 \ 1] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + [0] \quad (7.4.4)$$

where w_1 and w_2 are independent, unity white noise driving functions. Notice that v is zero and thus R^{-1} does not exist. Since we have a perfect measurement of the linear combination $x_1 + x_2$, let us transform to new states, y_1 and y_2 , according to

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (7.4.5)$$

or

$$y = Ax$$

Making this transformation leads to the new state and measurement models

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} \sqrt{2} & 0 \\ \sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad (7.4.6)$$

$$[z] = [0 \ 1] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + [0] \quad (7.4.7)$$

Note that z is now a perfect estimate of y_2 alone, so that we only need to worry about estimating y_1 . Thus, the order of the problem is reduced from 2 to 1.

Consider next the y_1 state equation

$$\dot{y}_1 = -y_1 + \sqrt{2}w_1 \quad (7.4.8)$$

This is in the correct form as it is. [Usually y_2 would also appear, and if so, it would be treated as a *known* driving function because it is equal to $z(t)$.]

Next, we need a measurement relationship of the proper form, that is, there must be nontrivial additive white noise and some linear connection to y_1 (and y_1 alone). In this case, since z has zero additive white noise, we can consider its derivative \dot{z} as also being available as a measurement.

$$\dot{z} = \dot{y}_2 = -y_2 + \sqrt{2}w_1 + \sqrt{2}w_2 \quad (7.4.9)$$

Observe that \dot{z} contains additive white noise as required. However, it also involves y_2 . In order to eliminate y_2 , we note

$$z = y_2 \quad (7.4.10)$$

Now, add z and \dot{z} to eliminate y_2 . This leads to

$$\dot{z} + z = 0 \cdot y_1 + \sqrt{2}(w_1 + w_2) \quad (7.4.11)$$

We now have a linear connection between $\dot{z} + z$ and y_1 with additive white noise, which is in the correct form. In the new problem of estimating y_1 , we then have

$$\begin{aligned} F &= -1 \\ G &= \sqrt{2} \\ w &= w_1 \\ Q &= 1 \\ H &= 0 \\ v &= \sqrt{2}(w_1 + w_2) \end{aligned} \quad (7.4.12)$$

and $\dot{z} + z$ plays the role of the "measurement." (After all, if z is known, so is \dot{z} .)

Note that w and v are correlated, so we must use the correlated form of Kalman filter equations given in Section 7.3. Using the notation of Section 7.3, we then have the additional parameters

$$\begin{aligned} C &= \sqrt{2} \\ R &= 4 \\ G(Q - CR^{-1}C^T)G^T &= 1 \\ D &= GCR^{-1} = \frac{1}{2} \end{aligned} \quad (7.4.13)$$

Therefore, the optimal filter is given by the differential equation

$$\dot{\hat{y}}_1 = -\hat{y}_1 + \frac{1}{2}(\dot{z} + z) - 0 \cdot \hat{y}_1 \quad (7.4.14)$$

This will be recognized as the equivalent transfer function relationship

$$\frac{\hat{y}_1(s)}{z(s)} = \frac{\frac{1}{2}(s+1)}{(s+1)} = \frac{1}{2} \quad (7.4.15)$$

This checks with Wiener filter theory, that is,

$$\hat{y}_1 = \frac{1}{2}z \quad (7.4.16)$$

We now have optimal estimates of y_1 and y_2 . All we have to do is transform back to the x domain to get optimal estimates of x_1 and x_2 . This yields

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2}z \\ z \end{bmatrix} = \begin{bmatrix} \frac{1}{2}z \\ \frac{1}{2}z \end{bmatrix} \quad (7.4.17)$$

Thus, the optimal estimates of both x_1 and x_2 check with the results of the Wiener theory. ■

The procedure for dealing with colored measurement noise may now be summarized as follows:

1. Decouple the quantities that are known perfectly from those that are not. Use a linear transformation and, as a matter of convenience, define the new state variables such that the perfectly known variables are at the "bottom" of the new state vector. The bottom elements should then have a one-to-one correspondence to the perfect measurements.
2. Consider a reduced state estimation problem where the state vector contains only the "top" elements of the new state vector. These are not known perfectly from the measurement information, so a nontrivial estimation problem exists relative to these variables. Note that the perfectly known bottom elements that appear on the right side of the state equation may be replaced with measurements z_1, z_2, \dots . They may then be grouped with the driving functions because they are known quantities. The reduced state equations must fit the general form

$$\dot{\hat{x}}_T = \mathbf{F}_T \hat{x}_T + (\text{known inputs}) + (\text{white noise inputs}) \quad (7.4.18)$$

where the subscript T is used to indicate top elements of the transformed state vector.

3. Rearrange the measurement equation to achieve the desired form

$$\mathbf{z}_T = \mathbf{H}_T \hat{x}_T + (\text{white noise}) \quad (7.4.19)$$

Note that a nontrivial white noise component must be associated with each measurement element. Also, the number of elements in \mathbf{z}_T must be the same as for the original \mathbf{z} vector in order that no measurement information is

ignored in estimating \mathbf{x}_T . The white noise components are brought in by repeated differentiation of the perfect \mathbf{z} elements until a white noise term appears. The new \mathbf{z}_T vector is then formed from appropriate linear combinations of the elements of \mathbf{z} and their derivatives. Sometimes considerable algebraic manipulation is needed to twist the equations into the form demanded by Eq. (7.4.19), but this can always be done.

4. Solve the reduced state estimation problem. Usually the filter equations for correlated \mathbf{w} and \mathbf{v} , as given in Section 7.3, will have to be used. The result will be a differential equation for the reduced state vector. This equation will normally contain derivatives of some of the measurements, as well as the measurements, as driving functions. The derivatives usually present no problems and may be left there as driving functions. If, however, one wishes to remove them, there are standard techniques in linear system theory for doing so [e.g., see Ogata (3) or Chen (4)]. The reduced state equation is usually solved (conceptually, at least) for either the transient or steady-state solution, whichever is desired. The end result is the best estimate of the reduced state vector (i.e., the imperfectly known elements) in the transformed domain.
5. Finally, append the perfectly known elements to the bottom of those estimated in step 4, and transform the total state estimate back to the original state space. This gives the optimal estimate for the original problem.

7.5 SUBOPTIMAL ERROR ANALYSIS

Just as in the discrete case, an error covariance equation can be derived that is applicable to suboptimal as well as optimal gain. The derivation is similar to that given in Section 7.1, and the relationships between \mathbf{Q} and \mathbf{Q}_k and \mathbf{R} and \mathbf{R}_k are given there. We begin with the projection equation for the discrete filter

$$\mathbf{P}_{k+1}^- = \phi_k \mathbf{P}_k \phi_k^T + \mathbf{Q}_k \quad (7.5.1)$$

We now replace \mathbf{P}_k with the general \mathbf{P} -update expression given by Eq. (5.4.11). This yields

$$\mathbf{P}_{k+1}^- = \phi_k [(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^- (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)^T + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^T] \phi_k^T + \mathbf{Q}_k \quad (7.5.2)$$

Next, the following substitutions are made in Eq. (7.5.2):

$$\begin{aligned} \phi_k &= \mathbf{I} + \mathbf{F} \Delta t \\ \mathbf{K}_k &= \mathbf{K} \Delta t \\ \mathbf{Q}_k &= \mathbf{G} \mathbf{Q} \mathbf{G}^T \Delta t \\ \mathbf{R}_k &= \mathbf{R} / \Delta t \end{aligned} \quad (7.5.3)$$

Then, after neglecting higher-order terms in Δt , the expression for \mathbf{P}_{k+1}^- becomes

$$\begin{aligned} \mathbf{P}_{k+1}^- &= \mathbf{P}_k^- - \mathbf{K} \mathbf{H}_k \mathbf{P}_k^- \Delta t - \mathbf{P}_k^- \mathbf{H}_k^T \mathbf{K}^T \Delta t + \mathbf{K} \mathbf{R} \mathbf{K}^T \Delta t + \mathbf{F} \mathbf{P}_k^- \Delta t \\ &\quad + \mathbf{P}_k^- \mathbf{F}^T \Delta t + \mathbf{G} \mathbf{Q} \mathbf{G}^T \Delta t \end{aligned} \quad (7.5.4)$$

Finally, forming the difference $\mathbf{P}_{k+1}^- - \mathbf{P}_k^-$, dividing by Δt , and passing to the limit leads to

$$\dot{\mathbf{P}} = \mathbf{F} \mathbf{P} + \mathbf{P} \mathbf{F}^T - \mathbf{K} \mathbf{H} \mathbf{P} - \mathbf{P} \mathbf{H}^T \mathbf{K}^T + \mathbf{K} \mathbf{R} \mathbf{K}^T + \mathbf{G} \mathbf{Q} \mathbf{G}^T \quad (7.5.5)$$

This equation may now be used with any gain \mathbf{K} and is useful in suboptimal error analysis. A word of caution is in order, though. Note "gain" means the \mathbf{K} coefficient in the differential equation

$$\dot{\mathbf{x}} = \mathbf{F} \mathbf{x} + \mathbf{K} (\mathbf{z} - \mathbf{H} \mathbf{x}) \quad (7.5.6)$$

and the equations describing the suboptimal filter under consideration must be put into the proper format before Eq. (7.5.5) can be applied [see Problem 7.3 for an example of the use of Eq. (7.5.5)].

7.6 RELATIONSHIP BETWEEN WIENER AND KALMAN FILTERS

It is appropriate here to pause and reflect on the connection between Wiener and Kalman filter theories. Both the Wiener and Kalman filters are minimum mean-square-error estimators, both require the same a priori information about the processes being estimated, and both yield identical estimates. We saw in Chapter 4 that the Wiener approach leads to an integral equation with the filter weighting function as the unknown. After solution of the integral equation, the weighting function then describes the relationship between input and output. On the other hand, the end result of continuous Kalman filter theory is a differential equation relating input and output. There must be equivalence, and books on linear system theory [e.g., Chen (4)] may be consulted for the details of how to get back and forth between the two descriptions. There are, however, certain subtleties about this particular problem that warrant some further amplification.

First, in the Wiener theory we related the filter response to the input via the superposition (convolution) integral, which was written in the form

$$\hat{x}(t) = \int_0^t g(\tau, t) z(t - \tau) d\tau \quad (7.6.1)$$

This form was convenient because the first argument of g has the significance of the "age" variable, and g tells us how the past values of the input are weighted

to yield the present value of the output (5). The second argument of g (i.e., t) simply appears as a parameter that may be considered fixed in the optimization process. Frequently, though, books on linear system theory [e.g., see Chen (4)] write the superposition integral in the form

$$\hat{x}(t) = \int_0^t h(t, \tau) z(\tau) d\tau \quad (7.6.2)$$

(This was mentioned before in Chapter 4.) When \hat{x} is written in the form of Eq. (7.6.2), $h(t, \tau)$ has a physical interpretation as the system response at time t to a unit impulse applied at time τ . By making a simple change of dummy variables in Eq. (7.6.1), we can obtain the relationship between g and h :

$$g(t - \tau, t) = h(t, \tau) \quad (7.6.3)$$

This is not a point of confusion in constant-parameter systems. It is, however, in time-variable systems, and we have to be watchful of this little detail in converting from a weighting-function description to a state-realization (differential equation) description.

In addition, the Wiener filter is always a single-output estimator. That is, we may have multiple inputs, but we always choose a single scalar process (usually called "signal") as the quantity being estimated. For example, consider the problem where we have one Markov process that we call signal s , and the additive noise consists of another Markov process n_1 plus a white component n_2 . In Wiener theory, n_1 gets lumped with n_2 as the corrupting noise and the estimator yields just an estimate of s . On the other hand, with Kalman filtering, one models both s and n_1 as state variables, and the filter estimates both s and n_1 simultaneously (even though n_1 may not be of interest). In effect, the Kalman filter does the work of two Wiener filters. Furthermore, the Kalman filter automatically provides information about the quality of the estimates (i.e., their mean-square errors) while doing the estimation. The Wiener filter does not provide this information, and one has to go to considerable extra effort to get mean-square error information. Thus, we have in the Kalman filter a group of estimators, all packaged into a single matrix algorithm. The price, of course, is extra computation effort. We often find in Kalman filtering that we are forced to carry along considerable excess baggage in order to obtain estimates of a few select quantities of interest.

Another subtlety that appears when comparing Wiener and Kalman methods has to do with the initial conditions. It may appear at first glance that we are allowed to choose \hat{x}_0 and P_0 as we wish in a Kalman filter, whereas no such explicit choice exists in the corresponding Wiener formulation. This apparent discrepancy exists because the Wiener filter output was written as a superposition integral, which tacitly implies zero initial conditions for the filter. This is justified, provided the input processes have zero mean and we have no prior information about the processes other than what was already assumed in modeling the various spectral functions. To get correspondence in the Kalman filter

we must choose \hat{x}_0 to be zero; and, having done this, P_0 must be the covariance of the process itself. Thus, we really do not have as much legitimate choice in the initial \hat{x}_0 and P_0 as it seems at first glance; at least this is true if we want the result to be optimal. The correspondence between the Kalman and Wiener solutions, properly initialized, is easily verified for simple cases and Problems 7.4 and 7.5 are intended to demonstrate this.

In summary, it is not proper to say that a Kalman filter is either better or worse than a Wiener filter, because both yield identical results once they are realized. However, in the realization, either for analysis purposes or on-line, the Kalman approach clearly has two distinct advantages:

1. With one common matrix formulation, we can accommodate a large class of estimation problems with relatively complicated process and measurement relationships.
2. The recursive feature of the Kalman filter makes it readily adaptable to computer solutions. This is certainly of considerable practical importance. Solutions of relatively complex estimation problems are often quite feasible using Kalman filtering methods, whereas the same problems may be completely intractable using Wiener methods.

Problems

7.1. Consider a scalar process $x(t)$ that may be thought of as the result of integrating Gaussian white noise with a spectral amplitude of 16 units. Let us say the integrator is "zeroed" at $t = 0$, and thus we know $x(0) = 0$. (This is a Wiener process.) Let us further say that we have a continuous noisy measurement of x where the measurement noise is white, Gaussian, independent of x , and has a spectral amplitude of 4 units.

- (a) Find the optimal Kalman filter for this situation. (Your solution may be left in the form of a differential equation, but all parameters of the differential equation are to be written out explicitly.)
- (b) Find the optimal filter transfer function and the corresponding rms error for the steady-state condition.

7.2. In the two-input Wiener problem described in Problem 4.17, there were two independent noisy measurements of the signal:

$$z_1 = s(t) + n_1(t)$$

$$z_2 = s(t) + n_2(t)$$

The autocorrelation functions for the signal and noises were given as

$$R_s(\tau) = e^{-|\tau|}$$

$$R_{n_1}(\tau) = e^{-2|\tau|}$$

$$R_{n_2}(\tau) = \delta(\tau)$$

and all three processes were independent of each other. Solve this same problem again for the steady-state transfer functions for the z_1 and z_2 inputs using Kalman filtering

methods. (*Hint:* One of the measurement noises is colored, so the procedure given in Section 7.4 should be helpful.)

7.3. The optimal estimator for a Markov signal plus additive white noise was given in Example 7.1. The gain worked out to be time variable, but it did approach a constant value of $(\sqrt{3} - 1)$ as $t \rightarrow \infty$. Investigate the effect of using this constant gain for all $t \geq 0$ relative to the filter rms error. Sketch plots of both the optimal and suboptimal rms errors for purposes of comparison.

7.4. Consider a stationary Gaussian Markov signal $x(t)$ whose autocorrelation function is

$$R_x(\tau) = 4e^{-|\tau|}$$

We make two noisy measurements of the signal, one at $t = 0$ and the other at $t = 1$, and we denote these as z_0 and z_1 . Each discrete measurement is known to have an error variance of unity, and the errors are statistically independent. We have no prior knowledge of the signal other than its autocorrelation function as stated above.

- (a) Using the methods of Section 4.8 (i.e., a weighting function approach), write an explicit expression for the optimal estimate of $x(1)$ in terms of z_0 and z_1 .
- (b) Repeat part (a) using discrete Kalman filtering and compare the result with that obtained in part (a). (*Note:* The Wiener and Kalman estimates should be identical.)
- (c) Let the initial conditions for a discrete Kalman filter be

$$\hat{x}_0^- = 0 \text{ and } P_0^- = \infty$$

Then work out the explicit expression for $\hat{x}(1)$ and note that it is different from those obtained in parts (a) and (b).

7.5. The same estimation problem was used in both Examples 4.5 and 7.1. Show that the two estimators yield identical results for all $t > 0$ as well as for the steady-state condition. (*Hint:* In the Wiener solution, $\hat{x}(t)$ can be written in terms of a convolution integral where the weighting function is known explicitly. First, convert the weighting function to impulsive-response form and then substitute the integral expression for $\hat{x}(t)$ into the differential equation describing the Kalman estimator and show that it satisfies the differential equation.)

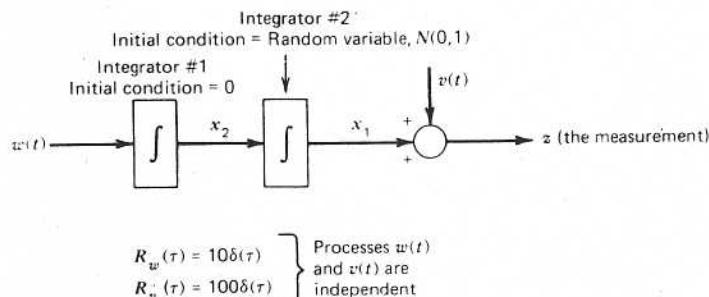
7.6. A certain noisy measurement is known to be of the form

$$z(t) = a_0 + n(t), \quad t \geq 0$$

where a_0 is an unknown random constant and $n(t)$ is white Gaussian noise with a spectral amplitude of A . Initially, at $t = 0$, all that is known about a_0 is that it has a zero-mean normal distribution with a variance σ^2 . We wish to estimate a_0 on a "running time" basis beginning at $t = 0$. Find the appropriate continuous Kalman filter for estimating a_0 and sketch a block diagram for the filter. (*Note:* This is the same estimation problem as that of Problem 4.12. However, when using Kalman filtering, we do not need to consider a random constant as a limiting case of a Markov process.)

7.7. The resulting optimal filter for Problem 7.6 is described by a differential equation relating the input $z(t)$ to the output $\hat{x}(t)$. When the same estimation problem is solved with Wiener methods, the result is a weighting function relating $z(t)$ and $\hat{x}(t)$. Show the two results are equivalent.

7.8. Consider the two-state Gaussian random process shown in the block diagram below. Clearly, the measurement is of state variable #1, and it may be assumed that the measurement begins at $t = 0$. The initial condition on the first integrator (state variable #2) is zero, whereas the second integrator has an initial condition that is a zero-mean normal random variable with $\sigma^2 = 1$. Write the differential equation for the P matrix for the continuous Kalman filter for this situation. Be sure to specify the appropriate initial conditions for P , and also be sure to specify all elements of the matrix parameters (such as F , R , and Q) of the differential equation. You do not need to solve the differential equation.



Problem 7.8

7.9. Modification of the discrete Kalman filter equations to include a deterministic input was discussed in Section 6.9. Show that the corresponding modification in the continuous case is accomplished with the addition of Bu as a driving function in the $\dot{\hat{x}}$ differential equation. Specifically, assume the process equation is of the form

$$\dot{\hat{x}} = F\hat{x} + Gw + Bu \quad (P7.9.1)$$

where Bu is the deterministic input. Then show that the corresponding differential equation for the estimate is

$$\dot{\hat{x}} = F\hat{x} + K(z - H\hat{x}) + Bu \quad (P7.9.2)$$

References Cited in Chapter 7

1. R. E. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction," *Trans. of the ASME, Jour. of Basic Engr.*, 83: 95-108 (1961).
2. A. E. Bryson, Jr., and D. E. Johansen, "Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise," *IEEE Trans. on Automatic Control*, AC-10: No. 1, 4-10 (Jan. 1965).
3. K. Ogata, *State Space Analysis of Control Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1967.
4. C. T. Chen, *Introduction to Linear System Theory*, New York: Holt, Rinehart, and Winston, 1970.

5. R. G. Brown and J. W. Nilsson, *Introduction to Linear Systems Analysis*, New York: Wiley, 1962.

Additional References on Continuous Kalman Filtering

6. A. Gelb (Ed.), *Applied Optimal Estimation*, Cambridge, Mass.: MIT Press, 1974.
7. A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communications and Control*, New York: McGraw-Hill, 1971.
8. J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*, New York: McGraw-Hill, 1969.

CHAPTER 8

Discrete Smoothing and Prediction

The emphasis in the preceding chapters on Kalman filtering has been on the *filter* problem, that is, estimation of the process at the same time as the current measurement. We now wish to generalize the theory to include prediction and smoothing just as was done with the Wiener theory. It works out that discrete prediction is relatively easy, so it will be considered first. The smoothing problem is somewhat more difficult; but yet it is still manageable and recursive solutions are available. We will see that the Kalman theory accommodates a larger class of smoothing-type problems than the Wiener theory, and thus it has more potential in applied work, especially in off-line data processing. We now proceed to the discrete prediction problem.

8.1 PREDICTION

Recall that in the discrete filter equations there is a projection step where the a posteriori estimate $\hat{\mathbf{x}}_k$ is projected ahead to yield the a priori estimate $\hat{\mathbf{x}}_{k+1}^-$. This is, of course, the best estimate of \mathbf{x} at t_{k+1} , given all the measurement data up through \mathbf{z}_k . Thus, this is one-step prediction. To predict N steps ahead we simply use the same projection equations (and justification) with an appropriate change in the ϕ and \mathbf{Q} matrices. Thus

$$\hat{\mathbf{x}}(k+N|k) = \phi_{k+N,k} \hat{\mathbf{x}}(k|k) \quad (8.1.1)$$

$$\mathbf{P}(k+N|k) = \phi_{k+N,k} \mathbf{P}(k|k) \phi_{k+N,k}^T + \mathbf{Q}_{k+N,k} \quad (8.1.2)$$

where

$\hat{\mathbf{x}}(k+N|k)$ = estimate of \mathbf{x} at t_{k+N} given measurement data through \mathbf{z}_k

$\mathbf{P}(k+N|k)$ = error covariance associated with $\hat{\mathbf{x}}(k+N|k)$

$\phi_{k+N,k}$ = state transition matrix for the time interval from t_k to t_{k+N}

$\mathbf{Q}_{k+N,k}$ = covariance of the driven response for the time interval from t_k to t_{k+N}

$\hat{\mathbf{x}}(k|k)$, $\mathbf{P}(k|k)$ = filter estimate and its error covariance at t_k

A word of explanation about notation is in order here. In both smoothing and prediction we need to keep track of the estimate and measurement times

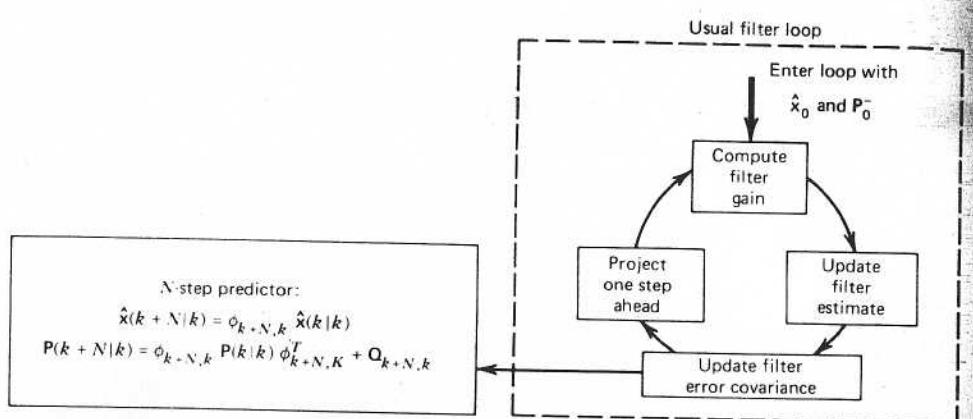


Figure 8.1. *N*-step prediction.

separately. This will be done with the parenthetical notation indicated above; for example, we read $\hat{x}(k+N|k)$ as “estimate of \mathbf{x} at time t_{k+N} given all the measurement data up through time t_k .” Our previous single-subscript notation then translates into:

$$\hat{x}_{\bar{k}} = \hat{x}(k|k-1)$$

$$P_{\bar{k}} = P(k|k-1)$$

$$\hat{x}_k = \hat{x}(k|k)$$

$$P_k = P(k|k)$$

The beauty of the explicit parenthetical notation is that prediction, filtering, and smoothing can all be accommodated within the same notational structure. The price, of course, is more writing.

Note that the expressions for $\hat{x}(k+N|k)$ and $P(k+N|k)$ involve the filter quantities $\hat{x}(k|k)$ and $P(k|k)$. Thus, the recursive filter equations must be solved as usual. The prediction of \mathbf{x} at t_{k+N} is an extra computation. This is summarized in Fig. 8.1. The equation for the predictor error covariance $P(k+N|k)$ is included in the predictor box in Fig. 8.1. This error covariance is not needed in the recursive filter loop, so this computation may be omitted if $P(k+N|k)$ is not of interest.

8.2 CLASSIFICATION OF SMOOTHING PROBLEMS

Smoothing seems to be inherently more difficult than either filtering or prediction. In the Wiener theory of Chapter 4, α is negative in Eq. (4.3.22) and this

causes a cusp to appear in the positive-time part of the shifted time function. (At least this is true for rational spectral functions—see Problem 8.1.) This cusp, in turn, leads to a considerably more complicated expression for the Fourier transform than occurs in the corresponding filter or prediction problems. Similarly, we will see that the recursive algorithms for smoothing are considerably more complicated than those for filtering and prediction.

Meditch (1) classifies smoothing into three categories:

1. *Fixed-interval smoothing*. Here the time interval of measurements (i.e., the data span) is fixed, and we seek optimal estimates at some, or perhaps all, interior points. This is the typical problem encountered when processing noisy measurement data off-line.
2. *Fixed-point smoothing*. In this case, we seek an estimate at a single fixed point in time, and we think of the measurements continuing on indefinitely ahead of the point of estimation. An example of this would be the estimation of initial conditions based on noisy trajectory observations after $t = 0$. In fixed-point smoothing there is no loss of generality in letting the fixed point be at the beginning of the data stream, because all prior measurements can be processed with the filter algorithm.
3. *Fixed-lag smoothing*. In this problem, we again envision the measurement information proceeding on indefinitely with the running time variable t , and we seek an optimal estimate of the process at a fixed length of time back in the past. Clearly, the Wiener problem with α negative is fixed-lag smoothing. It is of interest to note that the Wiener formulation will not accommodate either fixed-interval or fixed-point smoothing without using multiple sweeps through the same data with different values of α . This would be a most awkward way to process measurement data.

Obviously, the three smoothing problems are related, and it is not especially difficult to devise correct, but clumsy, solutions. Thus, the central problem is one of finding reasonably efficient recursive algorithms for each of the types of smoothing. This has been studied extensively since the early 1960s, and we will present only those solutions that are generally considered to be best from a computational viewpoint. We will not attempt to derive all the algorithms. The derivations are adequately documented in the references cited. (Also, see Problem 8.2.) We now proceed to look at algorithms for the three specified categories of smoothing.

8.3 DISCRETE FIXED-INTERVAL SMOOTHING

The algorithm to be presented here is due to Rauch, Tung, and Striebel (2.3), and its derivation is given in Meditch (1) as well as the referenced papers. Also,

a new simplified derivation is presented in Problem 8.2. In the interest of brevity, the algorithm will be subsequently referred to as the RTS algorithm. Consider a fixed-length interval containing $N + 1$ measurements. These will be indexed in ascending order $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_N$. The assumptions relative to the process and measurement models are the same as for the filter problem. The computational procedure for the RTS algorithm consists of a forward recursive sweep followed by a backward sweep. This is illustrated in Fig. 8.2. We enter the algorithm as usual at $k = 0$ with the initial conditions $\hat{\mathbf{x}}_0^-$ and \mathbf{P}_0^- . We then sweep forward using the conventional filter algorithm. With each step of the forward sweep, we must save the computed a priori and a posteriori estimates and their associated \mathbf{P} matrices. These are needed for the backward sweep. After completing the forward sweep, we begin the backward sweep with "initial" conditions $\hat{\mathbf{x}}(N|N)$ and $\mathbf{P}(N|N)$ obtained as the final computation in the forward sweep. With each step of the backward sweep the old filter estimate is updated to yield an improved smoothed estimate, which is based on all the measurement data. The recursive equations for the backward sweep are

$$\hat{\mathbf{x}}(k|N) = \hat{\mathbf{x}}(k|k) + \mathbf{A}(k)[\hat{\mathbf{x}}(k+1|N) - \hat{\mathbf{x}}(k+1|k)] \quad (8.3.1)$$

where the smoothing gain $\mathbf{A}(k)$ is given by

$$\mathbf{A}(k) = \mathbf{P}(k|k)\boldsymbol{\phi}^T(k+1,k)\mathbf{P}^{-1}(k+1|k) \quad (8.3.2)$$

and

$$k = N - 1, N - 2, \dots, 0$$

The error covariance matrix for the smoothed estimates is given by the recursive equation

$$\mathbf{P}(k|N) = \mathbf{P}(k|k) + \mathbf{A}(k)[\mathbf{P}(k+1|N) - \mathbf{P}(k+1|k)]\mathbf{A}^T(k) \quad (8.3.3)$$

It is of interest to note that the *smoothing* error covariance matrix is not needed for the computation of the estimates in the backward sweep. This is in contrast

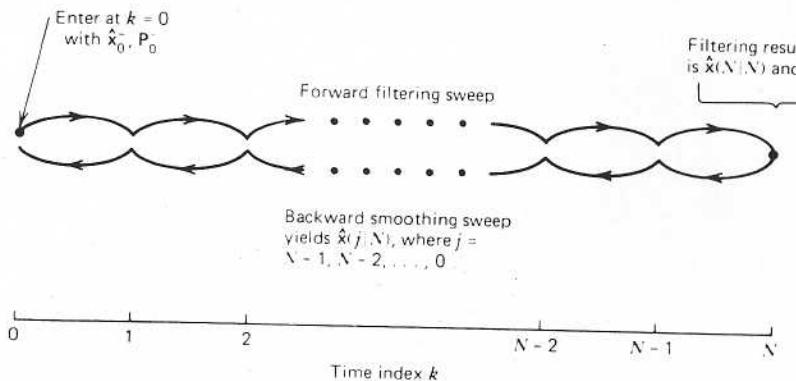


Figure 8.2. Procedure for fixed-interval smoothing.

to the situation in the filter (forward) sweep where the \mathbf{P} -matrix sequence is needed for the gain and associated estimate computations. An example illustrating the use of the RTS algorithm is now in order.

Example 8.1 Let us consider the same Gauss-Markov process used previously in Example 2, Section 5.5. Recall that the process has an autocorrelation function

$$R_x(\tau) = 1 \cdot e^{-|\tau|}$$

and we have a sequence of noisy measurements of this process taken every .02 sec. The measurement sequence begins at $t = 0$ and ends at $t = 1$ sec giving a total of 51 discrete measurements. The measurement errors are independent and have unity variance. A sample of this situation was simulated using Gaussian random numbers, and the filtering results were described previously in Section 5.5. We now continue this same example with a backward sweep and obtain smoothed estimates.

A partial listing of the results from the filtering solution are repeated in Table 8.1 for reference purposes. It can be seen from the error covariances that the filter has reached a steady-state condition by the end of the forward sweep.

We enter the backward sweep at the end point where $k = 50$. Here we have

$$\hat{\mathbf{x}}(50|50) = -.545 \quad (8.3.4)$$

$$\mathbf{P}(50|50) = .1653 \quad (8.3.5)$$

Since the filter solution at this point is conditioned on *all* the measurement data, it is also the smoothed estimate at $k = 50$. We are now ready to compute the smoothed estimate one step back at $k = 49$. From Eqs. (8.3.1) and (8.3.2) we have

$$\hat{\mathbf{x}}(49|50) = \hat{\mathbf{x}}(49|49) + \mathbf{A}(49)[\hat{\mathbf{x}}(50|50) - \hat{\mathbf{x}}(50|49)] \quad (8.3.6)$$

where

$$\mathbf{A}(49) = \mathbf{P}(49|49)\boldsymbol{\phi}^T(50|49) \quad (8.3.7)$$

Table 8.1 Filtering Results

k	$\hat{\mathbf{x}}(k k)$	$\mathbf{P}(k k)$	$\hat{\mathbf{x}}(k k-1)$	$\mathbf{P}(k k-1)$	$z(k)$
46	-.630	.1653	-.595	.1980	-.811
47	-.372	.1653	-.618	.1980	.871
48	-.521	.1653	-.364	.1980	-1.310
49	-.436	.1653	-.510	.1980	-.064
50	-.545	.1653	-.428	.1980	-1.138

Using the filtering results from Table 8.1, and noting that the transition matrix is $e^{-.02} \approx .98020$ leads to

$$A(49) = .8183$$

$$\hat{x}(49|50) = -.5317$$

This may now be repeated at $k = 48, 47, \dots$, etc., with the results

$$\hat{x}(48|50) = -.5388$$

$$\hat{x}(47|50) = -.5150$$

$$\hat{x}(46|50) = -.5457$$

.

etc.

The smoothing results for the entire 50 steps are summarized in Fig. 8.3 along with the filter results for comparison. It is evident from the plots that the smoothed plot is, in fact, smoother than the filter plot. This is as expected,

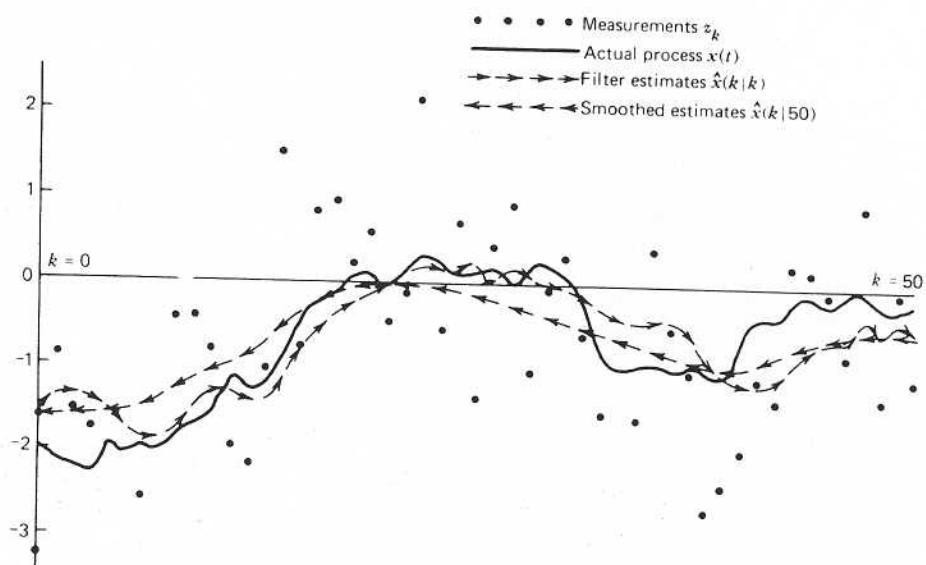


Figure 8.3. Smoothed and filter estimate plots for Example 8.1.

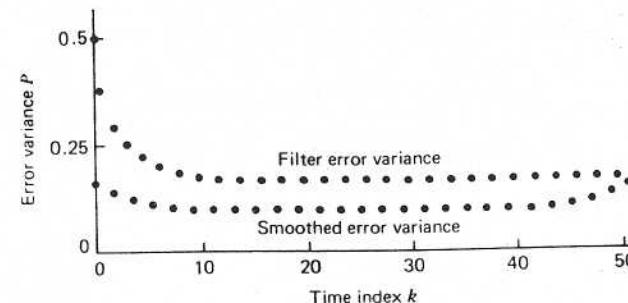


Figure 8.4. Error variances for filter and smoothed estimates of Example 8.1.

because the smoother uses both past and future data in the makeup of its estimate. The smoothed estimate is not conspicuously better than the filter estimate for this simulation. However, we should not expect to be able to draw firm conclusions from such a small sample.

It is also of interest in this example to compare the error variances of the filter and smoother estimates. These are shown in Fig. 8.4. Both the filter and smoother converge to a steady-state condition after about 15 steps. The steady-state values are

$$P_{\text{filter}} \approx .1653 \quad (\sigma_{\text{filter}} = .406)$$

$$P_{\text{smoother}} \approx .099 \quad (\sigma_{\text{smoother}} = .315)$$

Note that when the comparison is made on the basis of rms value, there is not a dramatic difference in the errors. In addition, the smoothed error curve is minimum in the middle and then gets larger as either end point is approached. This indicates that the best situation for estimation occurs in the interior region where there is an abundance of measurement data in either direction from the point of estimation. This is exactly what we would expect intuitively.

Before leaving this example, the steady-state behavior in the middle of the variance plot of Fig. 8.4 should be noted. This frequently happens when the data span is large and the process is stationary and observable. In the steady-state region, the filter and smoother gains and error covariance matrices are constant, and considerable computational simplification occurs, both in terms of the number of arithmetic operations and the amount of storage required for the algorithm. Thus, sometimes a problem that appears to be quite formidable at first glance works out to be feasible because a large amount of the measurement data can be processed in the steady-state region.

For future reference purposes, a partial listing of the fixed-interval smoothing solution is given in Table 8.2.

Table 8.2 Fixed-Interval Smoothing Solution

k	$x(k 50)$	$P(k 50)$
0	-1.601	.1653
1	-1.569	.1434
2	-1.567	.1287
3	-1.568	.1189
4	-1.565	.1123
5	-1.556	.1079
.	.	.
.	.	.
25	-1.146	.0990
.	.	.
.	.	.
45	-1.567	.1079
46	-1.546	.1123
47	-1.515	.1189
48	-1.539	.1287
49	-1.532	.1434
50	-1.545	.1653

8.4 DISCRETE FIXED-POINT SMOOTHING

The fixed-point smoothing algorithm to be presented here is taken directly from Meditch (1). Alternative algorithms that, under certain circumstances, may be better computationally are also given by Meditch. In order to keep the discussion here as simple as possible, we present just one algorithm, which was chosen because of its simplicity and similarity to the fixed-interval algorithm. The algorithm is

$$\hat{x}(k|j) = \hat{x}(k|j-1) + \mathbf{B}(j)[\hat{x}(j|j) - \hat{x}(j|j-1)] \quad (8.4.1)$$

where

$$k \text{ is fixed} \quad (\text{usually } k = 0)$$

$$j = k + 1, k + 2, \text{ etc.}$$

and

$$\mathbf{B}(j) = \prod_{i=k}^{j-1} \mathbf{A}(i) \quad (8.4.2)$$

$$\mathbf{A}(i) = \mathbf{P}(i|i)\phi^T(i+1|i)\mathbf{P}^{-1}(i+1|i) \quad (8.4.3)$$

(Note that the equation for \mathbf{A} is the same as for the fixed-interval algorithm.) The error covariance associated with the smoothed estimate is given by

$$\mathbf{P}(k|j) = \mathbf{P}(k|j-1) + \mathbf{B}(j)[\mathbf{P}(j|j) - \mathbf{P}(j|j-1)]\mathbf{B}^T(j) \quad (8.4.4)$$

Note that the filter estimates and their error covariances are needed in this algorithm, just as in the fixed-interval case. The computational procedure is summarized in Fig. 8.5. We enter the algorithm at the beginning of the data stream with the usual a priori \hat{x}_0^- and \mathbf{P}_0^- . (We have let $k = 0$.) These initial parameters may come from prior knowledge of the process in the event there are no prior measurements, or \hat{x}_0^- and \mathbf{P}_0^- may come from processing previous data via the filter algorithm. In either event, the first step is to update the a priori \hat{x}_0^- and \mathbf{P}_0^- with the measurement \mathbf{z}_0 and obtain $\hat{x}(0|0)$ and $\mathbf{P}(0|0)$. This is done with the usual filter equations and, of course, the result can be interpreted either as a filter estimate or the zero-stage smoothed estimate—they are one and the same. We next let $j = 1$, and we are ready to solve the one-stage smoothing problem using Eqs. (8.4.1) to (8.4.4). This may then be continued indefinitely, theoretically, at least. Of course, as with any recursive algorithm, round-off error may eventually lead to difficulties. We now look at an example of fixed-point smoothing.

Example 8.2 To illustrate the application of the fixed-point algorithm, we use the same process and simulated data used in Example 8.1. We let the fixed point be at $k = 0$ and the data for the first 20 steps are given in Table 8.3. Since the filtering results are needed for the solution of the smoothing problem, they are also listed in the table.

The first step is to compute $\hat{x}(0|0)$ and $\mathbf{P}(0|0)$. This was done in the previous filter solution with the result

$$\begin{aligned}\hat{x}(0|0) &= -1.610 \\ \mathbf{P}(0|0) &= .5\end{aligned}$$

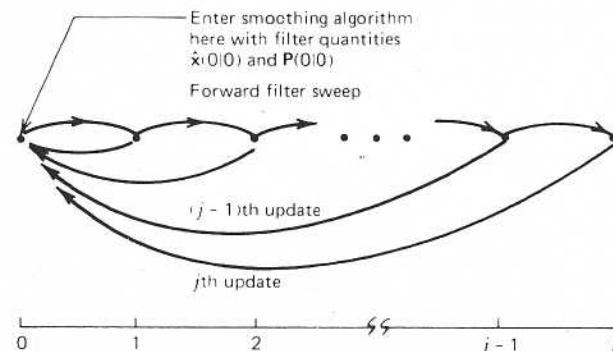


Figure 8.5. Procedure for fixed-point smoothing.

Table 8.3 Data for Fixed-Point Example

j	$\hat{x}(j j)$	$\hat{x}(j j-1)$	$A(j)$	$B(j)$	$\hat{x}(0 j)$	$P(j j)$	$P(j j-1)$	$P(0 j)$
0	-1.610		.9432		-1.610	.5		.5
1	-1.317	-1.578	.9114	.9432	-1.364	.3419	.5196	.3419
2	-1.350	-1.291	.8860	.8596	-1.414	.2689	.3677	.2689
3	-1.406	-1.323	.8661	.7616	-1.477	.2293	.2975	.2293
4	-1.454	-1.378	.8513	.6597	-1.528	.2060	.2595	.2060
5	-1.815	-1.425	.8411	.5616	-1.747	.1917	.2372	.1917
6	-1.911	-1.779	.8336	.4723	-1.809	.1826	.2234	.1826
7	-1.879	-1.873	.8287	.3937	-1.812	.1767	.2147	.1767
8	-1.592	-1.842	.8255	.3263	-1.730	.1729	.2090	.1729
9	-1.365	-1.560	.8227	.2693	-1.677	.1703	.2053	.1703
10	-1.238	-1.338	.8214	.2216	-1.655	.1686	.2029	.1686
11	-1.338	-1.214	.8201	.1820	-1.678	.1675	.2012	.1675
12	-1.436	-1.307	.8199	.1493	-1.697	.1668	.2002	.1668
13	-1.346	-1.408	.8191	.1224	-1.689	.1663	.1994	.1663
14	-8.848	-1.320	.8189	.1002	-1.642	.1660	.1990	.1660
15	-8.821	-.831	.8186	.0821	-1.641	.1657	.1987	.1657
16	-.540	-.804	.8186	.0672	-1.623	.1656	.1984	.1656
17	-.290	-.529	.8185	.0550	-1.610	.1655	.1983	.1655
18	-.204	-.284	.8184	.0450	-1.607	.1654	.1982	.1654
19	-.077	-.200	.8184	.0368	-1.602	.1654	.1981	.1654
20	-.134	-.076		.0302	-1.604	.1653	.1981	.1653

We are now ready to find $\hat{x}(0|1)$ by letting $k = 0$ and $j = 1$ in Eqs. (8.4.1), (8.4.2), and (8.4.3). The results are

$$\begin{aligned} B(1) &= A(0) = P(0|0)\phi P^{-1}(1|0) \\ &= .9432 \end{aligned}$$

$$\begin{aligned} \hat{x}(0|1) &= \hat{x}(0|0) + B(1)[\hat{x}(1|1) - \hat{x}(1|0)] \\ &= -1.364 \end{aligned}$$

We now have the needed information to go on to the next step and compute $\hat{x}(0|2)$, and so on. Note that it is not necessary to compute the smoothed error covariance in order to find the smoothed estimates. It was computed for checking purposes, though, and is listed in Table 8.3. The solution has nearly converged after 20 steps, and the results are essentially the same as those obtained from the fixed-interval solution, that is,

$$\hat{x}(0|20) \approx x(0|50) \approx -1.60$$

$$P(0|20) \approx P(p|50) \approx .1653$$

This indicates that the correlation between the process and the measurement 20 steps away from the point of estimation is so weak that measurements beyond 20 steps are of little value in forming the smoothed estimate. ■

8.5 FIXED-LAG SMOOTHING

The fixed-lag smoothing problem was originally solved by Wiener in the 1940s. However, he only considered the stationary case, that is, the smoother was assumed to have the entire past history of the input available for weighting in its determination of the estimate. Of the three smoothing categories mentioned in Section 8.2, the fixed-lag problem is the most complicated. One reason for this is the start-up problem. For example, let us say we wish to estimate our process at a point three steps back from the current point of measurement. If we enter the problem with the first measurement at $t = 0$, we have only the one measurement on which to base the estimate, and there is a gap in the measurement sequence between the current time and the point of estimation. This gap becomes filled as time progresses; but we have to be watchful of this little detail for the first few steps.

Meditch (1) gives an algorithm for fixed-lag smoothing that is considerably more complicated than the algorithms for the other two smoothing categories. We take a somewhat simpler approach here. Knowing the RTS algorithm for fixed-interval smoothing, we can always do fixed-lag smoothing by first filtering up to the current measurement and then sweeping back a fixed number of steps with the RTS algorithm. If the number of backward steps is small, this is a simple and effective way of doing fixed-lag smoothing. If, however, the number of backward steps is large, this method becomes cumbersome and one should look for more efficient algorithms. To illustrate the sweeping back on a running time basis and the start-up problem, we return to the Gauss-Markov example previously considered in Examples 8.1 and 8.2.

Example 8.3 The needed filtering data for our Gauss-Markov example is given in Example 8.2. Let us say that we wish to estimate the process three steps back from the current point of measurement, and that the first measurement occurs at $k = 0$. Index k is the discrete running time variable. Our first estimate is then to be $\hat{x}(-3|0)$. Anticipating that we will use the Rauch backward sweep from $k = 0$ to $k = -3$, we see that we will need all the filter estimates for this interval. Now, even though we have no measurements prior to $k = 0$, we can still form a trivial forward filtering sweep beginning at $k = -3$. Recall that filter updating without the benefit of a measurement simply amounts to setting the a posterior \hat{x} and P equal to the a priori \hat{x}^- and P^- . The result of this trivial sweep is shown on the left in Table 8.4. We now have all the filtering data needed for the backward sweep. The desired estimate $\hat{x}(-3|0)$ is obtained from Eqs. (8.3.1) and (8.3.2), and the results are given on the right side of Table 8.4. Presumably, the only estimate of interest here is the last entry in the table, $\hat{x}(-3|0) = -1.516$.

Next, suppose we get a measurement at $k = 1$, and we wish the smoothed estimate three steps back at $k = -2$. The forward sweep that began at $k = -3$

Table 8.4 Forward and Backward Sweeps (Current Measurement Is at $k = 0$)

k	Forward Sweep				Backward Sweep		
	$\hat{x}(k k-1)$	$P(k k-1)$	$\hat{x}(k k)$	$P(k k)$	k	$A(k)$	$\hat{x}(k 0)$
-3	0	1	0	1	0	—	-1.610
-2	0	1	0	1	-1	.9802	-1.578
-1	0	1	0	1	-2	.9802	-1.547
0	0	1	-1.610	.5	-3	.9802	-1.516

Table 8.5 Forward and Backward Sweeps (Current Measurement Is at $k = 1$)

k	Forward Sweep				Backward Sweep		
	$\hat{x}(k k-1)$	$P(k k-1)$	$\hat{x}(k k)$	$P(k k)$	k	$A(k)$	$\hat{x}(k 1)$
-2	0	1	0	1	1	—	-1.317
-1	0	1	0	1	0	.9432	-1.364
0	0	1	-1.610	.5	-1	.9802	-1.337
1	-1.578	.5196	-1.317	.3419	-2	.9802	-1.310

Table 8.6 Forward and Backward Sweeps (Current Measurement Is at $k = 2$)

k	Forward Sweep				Backward Sweep		
	$\hat{x}(k k-1)$	$P(k k-1)$	$\hat{x}(k k)$	$P(k k)$	k	$A(k)$	$\hat{x}(k 2)$
-1	0	1	0	1	2	—	-1.350
0	0	1	-1.610	.5	1	.9114	-1.371
1	-1.578	.5196	-1.317	.3419	0	.9432	-1.415
2	-1.291	.3677	-1.350	.2689	-1	.9802	-1.387

Table 8.7 Forward and Backward Sweeps (Current Measurement Is at $k = 3$)

k	Forward Sweep				Backward Sweep		
	$\hat{x}(k k-1)$	$P(k k-1)$	$\hat{x}(k k)$	$P(k k)$	k	$A(k)$	$\hat{x}(k 3)$
0	0	1	-1.610	.5	3	—	-1.406
1	-1.578	.5196	-1.317	.3419	2	.8860	-1.423
2	-1.291	.3677	-1.350	.2689	1	.9114	-1.438
3	-1.323	.2975	-1.406	.2293	0	.9432	-1.478

may now be extended one step, and the results are shown at the left in Table 8.5. Note that the only new entries are in the bottom row. The results of the backward sweep are also given in Table 8.5, and all the entries here are new; that is, the backward sweep must be completely redone with each new measurement. Note that each time we increment the filter forward one step, the oldest estimate and its error covariance may be discarded, because we only need the four most recent filter results for the backward sweep. Thus, we have sort of a "sliding window" of data that we must store as time increments forward.

The procedure just described may now be repeated indefinitely. For reference purposes, the solutions for the next two steps are given in Tables 8.6 and 8.7. The procedure has been demonstrated with tables at each step for the purpose of clarity, but it should be recognized that much of the data in the tables is either repetitious or not necessary for succeeding steps. Thus, the actual programming of the running forward-plus-backward algorithm is not as complicated as would appear from the tabular listings. With each new measurement, it simply involves moving one step forward and M steps back, where M is the desired fixed lag. ■

8.6 FORWARD-BACKWARD FILTER APPROACH TO SMOOTHING

In 1969, D. C. Fraser and J. E. Potter presented a novel solution to the smoothing problem (4). With reference to the fixed-interval problem, their approach was to *filter* the measurement data from both ends to the interior point of interest, and then the two filter estimates were combined to obtain the optimal smoothed estimate. This is illustrated in Fig. 8.6. The equations for the continuous problem will be developed first following the derivation given in Gelb (5). We will then deviate from both Gelb and the Fraser-Potter paper and present a new simplified derivation for the discrete case.

The Continuous Problem

Suppose we have two independent, unbiased estimates of some parameter x . Call these \hat{x}_1 and \hat{x}_2 and let the corresponding estimation errors have variances σ_1^2 and σ_2^2 . We wish to form a new estimate \hat{x} as a linear combination of \hat{x}_1 and \hat{x}_2 , where \hat{x} is to have minimum mean-square error. Thus, we write \hat{x} as

$$\hat{x} = k_1 \hat{x}_1 + k_2 \hat{x}_2 \quad (8.6.1)$$

If the new estimate is to be unbiased,

$$k_1 + k_2 = 1 \quad (8.6.2)$$

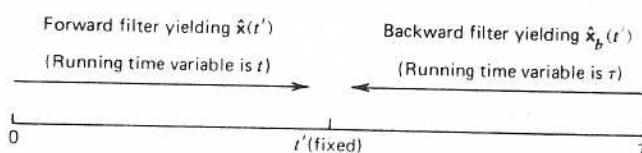


Figure 8.6. Forward-backward filtering. Smoothed estimate $\hat{x}(t'|T)$ is a linear combination of $\hat{x}(t')$ and $\hat{x}_b(t')$.

Thus, Eq. (8.6.1) may be rewritten as

$$\hat{x} = k_1 \hat{x}_1 + (1 - k_1) \hat{x}_2 \quad (8.6.3)$$

The mean-square error for \hat{x} is then

$$E[(x - \hat{x})^2] = E\{[x - k_1 \hat{x}_1 - (1 - k_1) \hat{x}_2]^2\} \quad (8.6.4)$$

or

$$E[e^2] = E[k_1(e_1 - e_2) + e_2]^2 \quad (8.6.5)$$

where e , e_1 , and e_2 are the errors in \hat{x} , \hat{x}_1 , and \hat{x}_2 , respectively. Equation (8.6.5) may now be differentiated with respect to k_1 and set equal to zero to find the optimal k_1 . Performing this operation and noting that e_1 and e_2 are independent yields

$$k_1 = \frac{E[e_2^2]}{E[e_1^2] + E[e_2^2]} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (8.6.6)$$

and

$$k_2 = \frac{E[e_1^2]}{E[e_1^2] + E[e_2^2]} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad (8.6.7)$$

The corresponding variance of the optimal estimate \hat{x} is obtained from Eq. (8.6.5) and is

$$\sigma^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (8.6.8)$$

Notice that the weighting factors k_1 and k_2 used for blending \hat{x}_1 and \hat{x}_2 are inversely proportional to the variances of the individual estimates. (Electrical engineers usually notice the obvious analogy between this and the division of electric current between two parallel resistors. Variance in the statistics problem plays the same role as resistance in the circuit problem.) We note in passing that Eqs. (8.6.6), (8.6.7), and (8.6.8) may also be written in terms of reciprocals as

$$\sigma^2 = [(\sigma_1^2)^{-1} + (\sigma_2^2)^{-1}]^{-1} \quad (8.6.9)$$

$$k_1 = \sigma^2 (\sigma_1^2)^{-1} \quad (8.6.10)$$

$$k_2 = \sigma^2 (\sigma_2^2)^{-1} \quad (8.6.11)$$

We now return to the forward-backward filter problem depicted in Fig. 8.6. The forward filter is the usual continuous Kalman filter as discussed in Chapter 7. Thus, the process model and filter equations are

Forward process model:

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{w} \quad \left. \begin{array}{l} \mathbf{w} \text{ and } \mathbf{v} \text{ are independent} \\ \mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v} \end{array} \right\} \text{white noise processes} \quad (8.6.12)$$

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v} \quad (8.6.13)$$

Forward filter equations:

$$\dot{\mathbf{x}} = \mathbf{F}\hat{\mathbf{x}} + \mathbf{P}\mathbf{H}^T \mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}) \quad (8.6.14)$$

$$\dot{\mathbf{P}} = \mathbf{FP} + \mathbf{P}\mathbf{F}^T - \mathbf{P}\mathbf{H}^T \mathbf{R}^{-1} \mathbf{HP} + \mathbf{G}\mathbf{Q}\mathbf{G}^T \quad (8.6.15)$$

Initial conditions:

$$\hat{\mathbf{x}}(0) = \mathbf{0}, \quad \mathbf{P}(0) = E[\mathbf{x}(0)\mathbf{x}^T(0)] \quad (8.6.16)$$

The resulting forward estimate and its associated error covariance will be denoted without subscripts as $\hat{\mathbf{x}}(t')$ and $\hat{\mathbf{P}}(t')$.

For the backward filter, it is convenient to define a new running time variable τ which proceeds backward in time. The backward process model is then obtained by replacing the time derivative in Eq. (8.6.12) with $-d/d\tau$. This leads to the process equation.

Backward process model:

$$\frac{d\mathbf{x}}{d\tau} = -\mathbf{Fx} - \mathbf{Gw} \quad (8.6.17)$$

The corresponding filter equations are then obtained by replacing \mathbf{F} and \mathbf{G} in Eqs. (8.6.14) and (8.6.15) with $-\mathbf{F}$ and $-\mathbf{G}$.

Backward filter equations:

$$\frac{d\hat{\mathbf{x}}_b}{d\tau} = -\mathbf{F}\hat{\mathbf{x}}_b + \mathbf{P}\mathbf{H}^T \mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}_b) \quad (8.6.18)$$

$$\frac{d\mathbf{P}_b}{d\tau} = -\mathbf{FP}_b - \mathbf{P}_b \mathbf{F}^T - \mathbf{P}_b \mathbf{H}^T \mathbf{R}^{-1} \mathbf{HP}_b + \mathbf{G}\mathbf{Q}\mathbf{G}^T \quad (8.6.19)$$

The boundary conditions for the backward filter equations are awkward, to say the least. Note that we have already used our a priori knowledge about the process in the forward filter. Thus, this information must not be used again in the backward filter. (If it were used in both filters, the two estimates, $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_b$, would not be independent when they meet at the interior point t' .) This forces us to demand that $\mathbf{P}_b(0) = \infty$. Note that $\tau = 0$ corresponds to $t = T$. The corresponding initial estimate of $\hat{\mathbf{x}}_b$ is arbitrary and immaterial, because the initial $\hat{\mathbf{x}}_b$ is given zero weight due to the infinite initial \mathbf{P}_b . Both Gelb and Fraser-Potter give a procedure for avoiding the "infinite- P " problem. Their procedure involves propagating \mathbf{P}_b^{-1} rather than \mathbf{P}_b . The boundary conditions, while awkward in the continuous problem, do not present similar problems in the discrete

counterpart, so we will be content here to beg the issue in the continuous problem and simply say that the initial conditions are awkward. (See Problem 8.6 for a further discussion of this.)

Assume now that forward and backward filters have been solved for $\hat{\mathbf{x}}$, \mathbf{P} and $\hat{\mathbf{x}}_b$, \mathbf{P}_b at the meeting point t' . The appropriate blend of $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_b$ and its error covariance are then obtained from Eqs. (8.6.1) and (8.6.9) to (8.6.11) by simply replacing scalar variances with covariance matrices. The final result is then

$$\mathbf{P}(t'|T) = (\mathbf{P}^{-1} + \mathbf{P}_b^{-1})^{-1} \quad (8.6.20)$$

$$\hat{\mathbf{x}}(t'|T) = \mathbf{P}(t'|T)[\mathbf{P}^{-1}\hat{\mathbf{x}} + \mathbf{P}_b^{-1}\hat{\mathbf{x}}_b] \quad (8.6.21)$$

The Discrete Problem

When experimental measurements are recorded for later processing off-line, they are usually sampled and recorded in digital form. Thus, the discrete smoothing problem is often of more practical interest than its continuous counterpart. It can be seen that the Fraser-Potter forward-backward filter approach has considerable potential for computational savings, as compared with the RTS algorithm. First, by filtering from both ends to the middle, there is no need to store any of the intermediate calculations along the way. The only ones needed for calculating the smoothed estimate are those obtained at the meeting point. Also, filtering does not involve inverting a \mathbf{P} matrix with each step, as is required on the backward sweep of the RTS algorithm. Since the \mathbf{P} matrix often has a much higher dimension than \mathbf{R} , having to form \mathbf{P}^{-1} with each step is a time-consuming computation. Thus, a discrete version of the Fraser-Potter approach is attractive and will now be pursued further.

The discrete forward filter has been discussed in some detail in Chapters 5 and 6, and there is no change in either the model or the recursive equations relative to the forward part of the estimator. As before, the a priori knowledge about the process will be incorporated into the forward filter and not the backward one. It is important to note, though, that the discrete forward filter equations are *not* approximations of the continuous filter equations. They are exact in their own right, provided that ϕ_k , \mathbf{Q}_k , \mathbf{H}_k , and \mathbf{R}_k are exact. We now seek a similar exact set of equations for the backward filter.

Once we have obtained an a priori estimate and its \mathbf{P} matrix at any point in the backward process, the updating is the same for the forward filter. This should be apparent from the derivation given in Chapter 5, because there is nothing in the derivation that depends on how we got the a priori estimate and its associated error covariance. Thus, the only different aspect in the backward procedure has to do with the projection of \mathbf{x}_b and \mathbf{P}_b from one step to the next. The projection equations can be obtained from the continuous backward filter equations, Eqs. (8.6.18) and (8.6.19). We do this by thinking of having arrived at some point i in the backward process, and then we proceed through the $(i, i+1)$ interval without benefit of measurements. Saying that we have no mea-

surements is the same as letting $\mathbf{R}^{-1} \rightarrow 0$, which indicates worthless measurements. If we let $\mathbf{R}^{-1} = 0$ in Eqs. (8.6.18) and (8.6.19), they reduce to

$$\frac{d\hat{\mathbf{x}}_b}{d\tau} = -\mathbf{F}\hat{\mathbf{x}}_b \quad (8.6.22)$$

$$\frac{d\mathbf{P}_b}{d\tau} = -\mathbf{FP}_b - \mathbf{P}_b\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T \quad (8.6.23)$$

These equations must now be solved for the $(i, i+1)$ interval, subject to the initial conditions $\hat{\mathbf{x}}_b(\tau_i)$ and $\mathbf{P}_b(\tau_i)$. (We assume here that i increments in unit steps in the same direction as τ .) The solution of Eq. (8.6.22) is obtained from linear system theory, just as in the case of the forward filter. In the interest of simplicity, we assume \mathbf{F} , \mathbf{G} , and \mathbf{Q} are constant over the interval. Then

$$\hat{\mathbf{x}}_b^*(\tau_{i+1}) = e^{-\mathbf{F}\Delta\tau}\hat{\mathbf{x}}_b(\tau_i) \quad (8.6.24)$$

or

$$\hat{\mathbf{x}}_{bi}^* = \phi^{-1}(\Delta\tau)\hat{\mathbf{x}}_{bi} \quad (8.6.25)$$

where the asterisk * indicates "a priori" and $\phi^{-1}(\Delta\tau)$ is the inverse of the forward transition matrix for the interval. This is exactly what we would expect intuitively.

Since the backward projection of \mathbf{P}_b is not as intuitively obvious as that for $\hat{\mathbf{x}}_b$, we will rely on the formal solution of Eq. (8.6.23) and not try to justify the answer intuitively. We will show that if \mathbf{P}_b is projected according to

$$\mathbf{P}_{bi+1}^* = \phi^{-1}(\Delta\tau)[\mathbf{P}_{bi} + \mathbf{Q}_i]\phi^{-1}(\Delta\tau)^T \quad (8.6.26)$$

the differential equation is satisfied. Let the right end of the projection interval be fixed at $\tau = \tau_0$. The left end is then the negatively running time variable τ , and $\tau > \tau_0$. We speculate that

$$\mathbf{P}_b(\tau) = \phi^{-1}(\tau - \tau_0)[\mathbf{P}_b(\tau_0) + \mathbf{Q}_i(\tau)]\phi^{-1}(\tau - \tau_0)^T \quad (8.6.27)$$

will satisfy Eq. (8.6.23). Note that the discrete \mathbf{Q}_i is written as a function of τ , because it varies as the interval increases. Differentiation of the assumed solution, Eq. (8.6.27), leads to

$$\frac{d\mathbf{P}_b}{d\tau} = \frac{d\phi^{-1}}{d\tau}[\mathbf{P}_b(\tau_0) + \mathbf{Q}_i]\phi^{-1} + \phi^{-1}\frac{d\mathbf{Q}_i}{d\tau}\phi^{-1} + \phi^{-1}[\mathbf{P}_b(\tau_0) + \mathbf{Q}_i]\frac{d\phi^{-1}}{d\tau}. \quad (8.6.28)$$

where the parenthetical dependence has been omitted in places to save writing. We first note from linear system theory that

$$\frac{d\phi^{-1}}{d\tau} = -\mathbf{F}\phi^{-1} \quad (8.6.29)$$

Thus, the first and third terms of Eq. (8.6.28) combine to yield $(-\mathbf{FP}_b - \mathbf{P}_b\mathbf{F}^T)$, which is identical to the first two terms on the right side of Eq. (8.6.23).

Note that the “measurement” in the linear model is the actual measurement less that predicted by the nominal trajectory in the absence of noise. Also the equivalent \mathbf{F} and \mathbf{H} matrices are obtained by evaluating the partial derivative matrices [Eqs. (9.1.8)] along the *nominal* trajectory. Example 9.1 illustrates this linearization procedure.

Example 9.1 This example is taken from Sorenson (4) and is a classic example of linearization of a nonlinear problem. Consider a near-earth space vehicle in a nearly circular orbit. It is desired to estimate the vehicle’s position and velocity on the basis of a sequence of angular measurements made with a horizon sensor. With reference to Fig. 9.1, the horizon sensor is capable of measuring:

1. The angle γ between the earth’s horizon and the local vertical.
2. The angle α between the local vertical and a known reference line (say to a celestial object).

In the interest of simplicity, we assume all motion and measurements to be within a plane as shown in Fig. 9.1. Thus the motion of the vehicle can be described with the usual polar coordinates r and θ .

The equations of motion for the space vehicle may be obtained from either Newtonian or Lagrangian mechanics. They are (see Section 2-10, Ref. 5):

$$\ddot{r} - r\dot{\theta}^2 + \frac{K}{r^2} = w_r(t) \quad (9.1.12)$$

$$r\ddot{\theta} + 2\dot{r}\dot{\theta} = w_\theta(t) \quad (9.1.13)$$

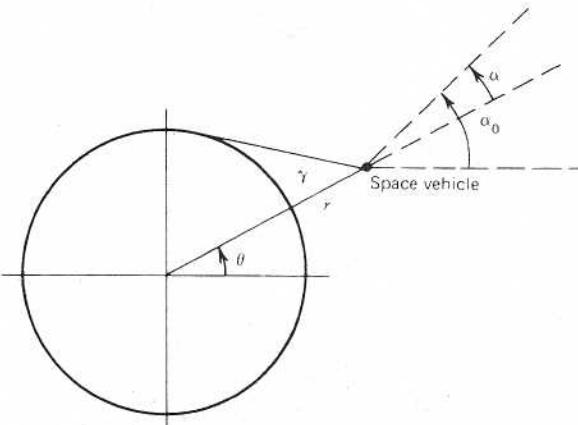


Figure 9.1. Coordinates for space vehicle example.

where K is a constant proportional to the universal gravitational constant, and w_r and w_θ are small random forcing functions in the r and θ directions (due mainly to gravitational anomalies unaccounted for in the K/r^2 term). We assume that the forcing functions are independent, white Gaussian random processes with known spectral amplitudes.

Let us first consider the nominal circular orbit. By direct substitution into Eqs. (9.1.12) and (9.1.13), it can be verified that

$$r^* = R_0 \quad (\text{a constant radius}) \quad (9.1.14)$$

$$\theta^* = \omega_0 t \quad \left(\omega_0 = \sqrt{\frac{K}{R_0^3}} \right) \quad (9.1.15)$$

will satisfy the undriven differential equations. Thus this is the reference trajectory.

At this point we could rewrite the differential equations of motion in state form and form the partial derivative matrix indicated in Eq. (9.1.8). This could then be evaluated along the reference trajectory to obtain the equivalent \mathbf{F} matrix for the linearized system (see Problem 9.1). In this example, though, the reference trajectory is especially simple, and it is easier to substitute $r^* + \Delta r$ and $\theta^* + \Delta\theta$ into Eqs. (9.1.12) and (9.1.13) before writing the equations in state form. Proceeding along this line yields

$$(\ddot{r}^* + \Delta\ddot{r}) - (r^* + \Delta r)(\dot{\theta}^* + \Delta\dot{\theta})^2 + \frac{K}{(r^* + \Delta r)^2} = w_r(t) \quad (9.1.16)$$

$$(r^* + \Delta r)(\ddot{\theta}^* + \Delta\ddot{\theta}) + 2(\dot{r}^* + \Delta\dot{r})(\dot{\theta}^* + \Delta\dot{\theta}) = w_\theta(t) \quad (9.1.17)$$

Now, taking advantage of Eqs. (9.1.14) and (9.1.15) and retaining only first-order terms in Δr and $\Delta\theta$ leads to

$$\Delta\ddot{r} - R_0\omega_0^2 - \omega_0^2\Delta r - 2R_0\omega_0\Delta\dot{\theta} + \frac{K}{R_0^2 + 2R_0\Delta r} = w_r(t) \quad (9.1.18)$$

$$R_0\Delta\ddot{\theta} + 2\omega_0\Delta\dot{r} = w_\theta(t) \quad (9.1.19)$$

We may now approximate $K/(R_0^2 + 2R_0\Delta r)$ as $K(1 - 2\Delta r/R_0)/R_0^2$. We also note that $K/R_0^2 = R_0\omega_0^2$. Equations (9.1.18) and (9.1.19) then reduce to

$$\Delta\ddot{r} - 3\omega_0^2\Delta r - 2R_0\omega_0\Delta\dot{\theta} = w_r(t) \quad (9.1.20)$$

$$\Delta\ddot{\theta} + 2\frac{\omega_0}{R_0}\Delta\dot{r} = w_\theta(t)/R_0 \quad (9.1.21)$$

These equations are linear and are ready to be put into the standard state-variable form. Let the state variables be

$$x_1 = \Delta r, \quad x_2 = \Delta\dot{r}, \quad x_3 = \Delta\theta, \quad x_4 = \Delta\dot{\theta} \quad (9.1.22)$$

Looking next at the $\phi^{-1}(d\mathbf{Q}_i/d\tau)\phi^{-1T}$ term in Eq. (8.6.28), we see that we need to form the derivative of \mathbf{Q}_i . Equation (5.3.6) is useful here. Let $t_k = T - \tau$, $t_{k+1} = T - \tau_0$, and note that $E[\mathbf{w}(u)\mathbf{w}^T(v)] = \mathbf{Q} \delta(u - v)$. After performing the inner integration we have

$$\mathbf{Q}_i = \int_{T-\tau}^{T-\tau_0} \phi(T - \tau, v) \mathbf{G} \mathbf{Q} \mathbf{G}^T \phi^T(T - \tau, v) dv \quad (8.6.30)$$

Differentiating this with respect to τ leads to

$$\frac{d\mathbf{Q}_i}{d\tau} = \phi(\tau - \tau_0) \mathbf{G} \mathbf{Q} \mathbf{G}^T \phi^T(\tau - \tau_0) \quad (8.6.31)$$

This may now be pre- and post-multiplied by $\phi^{-1}(\tau - \tau_0)$ and $\phi^{-1T}(\tau - \tau_0)$, as indicated in Eq. (8.6.28), and the result is obviously $\mathbf{G} \mathbf{Q} \mathbf{G}^T$, which equals the third term on the right side of Eq. (8.6.23). Thus, the discrete covariance projection given by Eq. (8.6.26) is seen to satisfy the continuous differential equation for \mathbf{P}_b with \mathbf{R}^{-1} set equal to zero. It is important to note that the ϕ and \mathbf{Q}_i in the projection equations are computed from the *forward* process equations. That is, they are the same parameters that would have been used in the forward filter for the interval, had the forward recursive steps been continued up through that interval of time.

The boundary conditions for the backward filter are awkward, but not impossible. One way of implementing the infinite initial \mathbf{P} matrix in off-line processing is to make all terms along its major diagonal very large, say about 10 orders of magnitude larger than their initial counterparts in the forward filter. The off-diagonal terms may then be set equal to zero. For all practical purposes, this completely deweights the prior information about the process relative to the weight that it is given in the forward filter. This is not a very elegant approach, but it is an effective, practical solution in many applications. In simple models where the order of the state vector is small, the boundary-condition problem may be handled analytically by using the alternative Kalman filter algorithm for the first step in the backward filter. This is illustrated in the one-state example presented at the end of this section, and it is also discussed further in Problem 8.6. If all else fails, we can resort to propagating \mathbf{P} inverse as suggested in the Fraser-Potter paper cited earlier.

We assume now that the forward filter has been stopped ahead recursively to the estimation point, say k , and the end result is an a posteriori estimate $\hat{\mathbf{x}}_k$ and an associated \mathbf{P}_k . The backward filter steps backward from the end point N , and it stops at $k + 1$ where it assimilates the \mathbf{z}_{k+1} measurement. It then projects this estimate one more step to obtain an *a priori* estimate $\hat{\mathbf{x}}_{b1}^*$ and its associated \mathbf{P}_{b1}^* . It does not assimilate the \mathbf{z}_k measurement, because this has already been used in the forward filter. Finally, the forward and backward estimates are blended together in accordance with the equation

$$\hat{\mathbf{x}}(k|N) = \mathbf{P}(k|N)[\mathbf{P}_k^{-1}\hat{\mathbf{x}}_k + \mathbf{P}_{b1}^{*-1}\hat{\mathbf{x}}_{b1}^*] \quad (8.6.32)$$

where

$$\mathbf{P}(k|N) = [\mathbf{P}_k^{-1} + \mathbf{P}_{b1}^{*-1}]^{-1} \quad (8.6.33)$$

An example will now illustrate the procedure.

Example 8.4 Again, we consider the same Gauss-Markov process used for the previous examples of this chapter. Let us say we want to find the smoothed estimate at $k = 48$, that is, $\hat{\mathbf{x}}(48|50)$. We first forward-filter up through \mathbf{z}_{48} . This has already been done in previous examples and the results are given in Table 8.1. The pertinent forward-filter results are

$$\hat{\mathbf{x}}_{48} = -.521$$

$$P_{48} = .165$$

We now need to generate a similar estimate at $k = 48$ with the backward filter. We begin at end point where $k = 50$. The initial conditions for the backward filter are

$$\hat{\mathbf{x}}_{b0}^* = 0, \quad P_{b0}^* = \infty \quad (8.6.34)$$

where the asterisk indicates “*a priori*” in the backward filter, and the subscript indicates the backward index i , rather than k . We update the initial estimate with the alternative Kalman filter algorithm in order to avoid the P -equal-infinity problem. Thus we have at $k = 50$ ($i = 0$)

$$P_{b0}^{-1} = P_{b0}^{*-1} + H_0^T R_0^{-1} H_0 \quad (8.6.35)$$

$$= 0 + 1$$

or

$$P_{b0} = 1 \quad (8.6.36)$$

The gain is then

$$K_{b0} = P_{b0} H_0^T R_0^{-1} = 1 \quad (8.6.37)$$

and the updated estimate is

$$\hat{\mathbf{x}}_{b0} = \hat{\mathbf{x}}_{b0}^* + K_{b0}[\mathbf{z}_0 - H_0 \hat{\mathbf{x}}_{b0}^*] = \mathbf{z}_0 = -1.138 \quad (8.6.38)$$

(from $k = 50$ entry of Table 8.1)

Note that we could have taken a more heuristic approach here and simply said: “Initially, we knew nothing about the backward process at the end point, and hence we must accept the measurement at $k = 50$ at face value. The resulting estimation error is then just the measurement error.” (This philosophy can also be extended to the vector case as shown in Problem 8.6.)

Continuing the backward filter, we next project $\hat{\mathbf{x}}_k$ and P_k to $k = 49$ ($i = 1$) using Eqs. (8.6.25) and (8.6.26). The results are

$$\hat{\mathbf{x}}_{b1}^* = e^{-0.02} \hat{\mathbf{x}}_{b0} = e^{-0.02}(-1.138) = -1.161 \quad (8.6.39)$$

$$P_{b1}^* = e^{-0.02}(P_{b0} + Q_i) e^{-0.02} = (e^{-0.02})^2 [1 + .0392] = 1.0816 \quad (8.6.40)$$

Next, we compute the gain and update the a priori estimate using the regular Kalman filter algorithm:

$$K_{b1} = \frac{P_{b1}^*}{P_{b1}^* + 1} = .5196 \quad (8.6.41)$$

$$x_{b1} = \hat{x}_{b1}^* + K_{b1}(z_1 - \hat{x}_{b1}^*) = -.591 \quad (8.6.42)$$

(Note from Table 8.1 the measurement at $k = 49$ is $-.064$.) The a posteriori error covariance matrix is then

$$P_{b1} = (1 - K_{b1})P_{b1}^* = .5196 \quad (8.6.43)$$

Now we project \hat{x}_{b1} and P_{b1} to $k = 48$ ($i = 2$).

$$\hat{x}_{b2}^* = e^{-0.02}\hat{x}_{b1} = -.603 \quad (8.6.44)$$

$$P_{b2}^* = e^{-0.02}(P_{b1} + Q_i)e^{-0.02} = .5816 \quad (8.6.45)$$

The backward filter is stopped at this point, because the measurement at $k = 48$ has already been assimilated in the forward filter.

We are now ready to blend together the forward and backward estimates in accordance with Eqs. (8.6.32) and (8.6.33). The results are

$$P(48|50) = [(.1653)^{-1} + (.5816)^{-1}]^{-1} = .1287$$

and

$$\hat{x}(48|50) = P(48|50)[P_{48}^{-1}\hat{x}_{48} + P_{b2}^{-1}\hat{x}_{b2}^*] = -.539$$

Notice that these results are the same as those given in Table 8.2, which were obtained using the RTS algorithm. ■

Problems

- 8.1.** Consider the same signal and noise situation used in Example 4.3. We found there that the causal Wiener filter was a simple first-order low-pass filter. Using Eq. (4.3.22) find the stationary, fixed-lag smoothing solution for $\alpha = -1$. Recall that the signal and noise are independent with autocorrelation functions

$$R_s(\tau) = e^{-|\tau|}$$

$$R_n(\tau) = \delta(\tau)$$

The solution may be given as a transfer function, that is, find $G(s)$ rather than $g(t)$. Note that the smoothing solution is considerably more complicated than the corresponding filter solution.

- 8.2.** (a) Consider the one-step-back, fixed-interval smoothing problem. Its solution may be obtained from the usual Kalman filter equations by batching together the last two measurements \mathbf{z}_{N-1} and \mathbf{z}_N and considering them as one measurement occurring at $N - 1$. Note that \mathbf{z}_N can be linearly related to \mathbf{x}_{N-1} as follows:

$$\begin{aligned} \mathbf{z}_N &= \mathbf{H}_N \mathbf{x}_N + \mathbf{v}_N \\ &= \mathbf{H}_N(\phi_{N-1} \mathbf{x}_{N-1} + \mathbf{w}_{N-1}) + \mathbf{v}_N \\ &= (\mathbf{H}_N \phi_{N-1}) \mathbf{x}_{N-1} + (\mathbf{H}_N \mathbf{w}_{N-1} + \mathbf{v}_N) \end{aligned}$$

The batched measurement relationship at $N - 1$ is then

$$\begin{bmatrix} \mathbf{z}_{N-1} \\ \mathbf{z}_N \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{N-1} \\ \mathbf{H}_N \phi_{N-1} \end{bmatrix} \mathbf{x}_{N-1} + \begin{bmatrix} \mathbf{v}_{N-1} \\ \mathbf{H}_N \mathbf{w}_{N-1} + \mathbf{v}_N \end{bmatrix}$$

Note that the upper and lower components of the batched measurement can be processed sequentially because their errors are uncorrelated. Assume now that we have an a priori estimate $\hat{\mathbf{x}}(N-1|N-2)$ and its error covariance matrix $\mathbf{P}(N-1|N-2)$. Proceed to update the estimate by assimilating the two components of the measurement separately in two steps, and show the final result is the same as that obtained using the RTS algorithm. (Hint: The first sequential step yields $\hat{\mathbf{x}}(N-1|N-1)$. Next, show that the gain for the second sequential step is $\mathbf{P}(N-1|N-1)\phi_{N-1}^T \mathbf{P}(N|N-1)^{-1} \mathbf{K}_N$ where \mathbf{K}_N is the usual Kalman filter gain for the N th stage. Finally, replace $\hat{\mathbf{x}}(N|N)$ in the RTS formula with $[\hat{\mathbf{x}}(N|N-1) + \mathbf{K}_N(\mathbf{z}_N - \mathbf{H}_N \hat{\mathbf{x}}(N|N-1))]$ and show the equivalence.)

- (b) The exercise of part (a) can be generalized to justify the RTS algorithm for any interior point within the fixed interval from 0 to N . To do this, let the interior point be denoted k and batch together all subsequent measurements $\mathbf{z}_{k+1}, \mathbf{z}_{k+2}, \dots, \mathbf{z}_N$. Call the batched measurement \mathbf{y}_{k+1} , that is,

$$\mathbf{y}_{k+1} = \begin{bmatrix} \mathbf{z}_{k+1} \\ \mathbf{z}_{k+2} \\ \vdots \\ \vdots \\ \mathbf{z}_N \end{bmatrix}$$

We can now form a linear connection between \mathbf{x}_{k+1} and \mathbf{y}_{k+1} as

$$\mathbf{y}_{k+1} = \mathbf{M}_{k+1} \mathbf{x}_{k+1} + \mathbf{u}_{k+1}$$

The batched measurement \mathbf{y}_{k+1} now plays the same role as \mathbf{z}_N in part (a). (We do not actually have to write out \mathbf{M}_{k+1} and \mathbf{u}_{k+1} explicitly. We simply need to know that such a relationship exists.) We can now consider the interior-point smoothing problem in terms of an equivalent filter problem that terminates at $k + 1$. This is the same as the problem considered in (a) except for notation. Proceed through the steps of exercise (a) again with appropriate changes in notation and show that the generalization is valid.

- 8.3.** Consider a stationary, Gauss-Markov process with an autocorrelation function

$$R_t(\tau) = e^{-|\tau|}$$

Assume that we have two noisy measurements of this process that were made at $t = 0$ and $t = 1$. Call these z_0 and z_1 . The measurement errors associated with z_0 and z_1 are independent and have a variance of unity. We wish to find the optimal estimate of x at

$t = 0$, given z_0 and z_1 , that is, we desire $\hat{x}(0|1)$. Write an expression for $\hat{x}(0|1)$ explicitly in terms of the measurements z_0 and z_1 using:

- The RTS algorithm.
- The Fraser-Potter forward-backward filter method.
- The weighting function method (see Section 4.8).
- The fixed-point algorithm of Section 8.4.

8.4. Show that the continuous version of the RTS algorithm is as follows:

$$\begin{aligned}\dot{\hat{x}}(t|T) &= \mathbf{F}\hat{x}(t|T) + \mathbf{G}\mathbf{Q}\mathbf{G}^T\mathbf{P}^{-1}(t|t)[\hat{x}(t|T) - \hat{x}(t|t)] \\ \dot{\mathbf{P}}(t|T) &= [\mathbf{F} + \mathbf{G}\mathbf{Q}\mathbf{G}^T\mathbf{P}^{-1}(t|t)]\mathbf{P}(t|T) + \mathbf{P}(t|T)[\mathbf{F} + \mathbf{G}\mathbf{Q}\mathbf{G}^T\mathbf{P}^{-1}(t|t)]^T - \mathbf{G}\mathbf{Q}\mathbf{G}^T\end{aligned}$$

Boundary Conditions: $\hat{x}(T|T)$ and $\mathbf{P}(T|T)$ obtained from filter solution.

(Hint: Begin with the discrete RTS algorithm and then let the step size approach zero, just as was done in deriving the filter equations in Chapter 7. Recall that \mathbf{Q}_k in the discrete model approaches $\mathbf{G}\mathbf{Q}\mathbf{G}^T \Delta t$ for small Δt . Also note that $[\mathbf{I} - \mathbf{A}(k)]$ is of the order of Δt in the smoothing algorithm; thus, the gain $\mathbf{A}(k)$ approaches \mathbf{I} in the limit as $\Delta t \rightarrow 0$.)

8.5. Consider a Wiener process and measurement situation as follows:

$$\dot{x} = w(t), \quad x(0) = 0$$

$$z = x + v(t)$$

where $w(t)$ and $v(t)$ are independent Gaussian white noise processes with spectral amplitudes q and r , respectively.

(a) Assume a fixed interval T that is sufficiently large to allow the filter solution to reach a steady-state condition on the forward sweep. This then becomes the boundary condition for the backward sweep. Using the continuous RTS algorithm given in Problem 8.4, show that $P(t|T)$ for the steady-state and terminal regions of the $[0, T]$ interval is approximated by

$$P(t|T) \approx \frac{\alpha}{2} [1 + e^{-2\beta(T-t)}]$$

where $\alpha = \sqrt{qr}$ and $\beta = \sqrt{q/r}$.

(b) Note that the solution of part (a) approaches $\alpha/2$ as $t \rightarrow 0$. This is obviously not compatible with the known a priori boundary condition for a Wiener process; that is, $P(0) = 0$. Explain this discrepancy.

8.6. Consider the two-state system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

where w_1 and w_2 are independent Gaussian white-noise inputs with unity-amplitude spectral functions. We have a sequence of scalar measurements of this process z_0, z_1, \dots, z_N that are related to the process by the equation

$$z_k = [1 \ 1] \mathbf{x}_k + v_k, \quad k = 0, 1, \dots, N$$

where v_k has unity variance. That is, we are only allowed to observe the sum of the state variables at each step and not their individual values. Suppose we want an estimate of

the process at some interior point, and we wish to get it using the Fraser-Potter forward-backward-filter method. In this case, the single measurement z_N does not provide enough information to yield a finite error-covariance estimate of x at $t = N$. Thus, we cannot start the backward filter quite as easily as was done in Example 8.4. In this problem, let the Δt interval be unity and show that the backward filter may be started at $t = N - 1$ by batching together z_N and z_{N-1} into an equivalent vector measurement at $t = N - 1$. The same technique used in Problem 8.2 will be helpful here. In effect, you need to show that the error covariance after assimilating z_N and z_{N-1} is finite and nonsingular, and that the estimate is the same as would be obtained by deterministic methods. (The extension of this technique to higher-order systems is fairly obvious, provided the system is observable. We simply batch together an appropriate number of measurements at the end of the interval and then solve for the system state, just as if the measurements were noisefree. This then becomes the initial state estimate for the backward filter, and we start the backward filtering an appropriate number of steps back from the end point.)

8.7. Table 8.3 summarizes the results of the fixed-point smoothing simulation of Example 8.2. Notice that the filter error variances listed under the column $P(j|j)$ are identical with the smoothing error variances $P(0|j)$. Give an intuitive explanation of this coincidence.

References Cited in Chapter 8

- J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*. New York: McGraw-Hill, 1969.
- H. E. Rauch, "Solutions to the Linear Smoothing Problem," *IEEE Trans. on Auto. Control*, AC-8: 371 (1963).
- H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum Likelihood Estimates of Linear Dynamic Systems," *AIAA J.*, 3: 1445 (1965).
- D. C. Fraser and J. E. Potter, "The Optimum Linear Smoother as a Combination of Two Optimum Linear Filters," *IEEE Trans. on Auto. Control*, AC-14: No. 4, 387 (August 1969).
- A. Gelb (Ed.), *Applied Optimal Estimation*. Cambridge, Mass.: MIT Press, 1974.

Additional References on Smoothing

- A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communications and Control*. New York: McGraw-Hill, 1971.
- B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall, 1979.

CHAPTER 9

Additional Topics on Kalman Filtering

Kalman's papers of the early 1960s (1,2) were recognized almost immediately as new and important contributions to least-squares filtering. As a result, there was a renewal of research interest in this area, and a flurry of papers expanding on Kalman's original work followed during the next decade or so. Kailath (3) gives an especially comprehensive bibliography of papers for this period. Research work in this area still continues (although perhaps at a somewhat reduced rate), and new applications and extensions continue to appear regularly in the technical literature. A few of the more significant extensions and related topics have been selected for comment here. The list is by no means comprehensive. However, there is a hierarchy of importance, and it is this author's recommendation that the reader begin with the first topic on linearization of nonlinear dynamics. The others may be studied in any desired order as time permits.

9.1 LINEARIZATION OF NONLINEAR PROCESS DYNAMICS

Some of the most successful applications of Kalman filtering have been in situations with nonlinear dynamics (e.g., vehicle trajectory determination). We now examine two basic ways of linearizing the problem dynamics. One is to linearize about some nominal trajectory in state space that does not depend on the measurement data. The resulting filter is usually referred to as simply a *linearized Kalman filter*. The other method is to linearize about a trajectory that is continually updated with the state estimates resulting from the measurements. When this is done, the filter is called an *extended Kalman filter*. A brief discussion of each will now be presented.

Linearized Kalman Filter

We begin by assuming the process to be estimated and the associated measurement relationship may be written in the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) + \mathbf{w}(t) \quad (9.1.1)$$

$$\mathbf{z} = \mathbf{h}(\mathbf{x}, t) + \mathbf{v}(t) \quad (9.1.2)$$

where \mathbf{f} and \mathbf{h} are known functions, \mathbf{u} is a deterministic forcing function, and \mathbf{w} and \mathbf{v} are independent white noise processes. Note that nonlinearity may enter into the problem either in the dynamics of the process or in the measurement relationship. Also, note that the forms of Eqs. (9.1.1) and (9.1.2) are somewhat restrictive in that \mathbf{w} and \mathbf{v} are assumed to be separate additive terms and are not included with the \mathbf{f} and \mathbf{h} terms. However, to do otherwise complicates the problem considerably, and thus we will stay with these restrictive forms.

Let us now assume that an approximate trajectory $\mathbf{x}^*(t)$ may be determined by some means. The actual trajectory $\mathbf{x}(t)$ may then be written as

$$\mathbf{x}(t) = \mathbf{x}^*(t) + \Delta\mathbf{x}(t) \quad (9.1.3)$$

Equations (9.1.1) and (9.1.2) then become

$$\dot{\mathbf{x}}^* + \Delta\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}^* + \Delta\mathbf{x}, \mathbf{u}, t) + \mathbf{w}(t) \quad (9.1.4)$$

$$\mathbf{z} = \mathbf{h}(\mathbf{x}^* + \Delta\mathbf{x}, t) + \mathbf{v}(t) \quad (9.1.5)$$

We now assume $\Delta\mathbf{x}$ is small and approximate the \mathbf{f} and \mathbf{h} functions with Taylor's series expansions, retaining only first-order terms. The result is

$$\dot{\mathbf{x}}^* + \Delta\dot{\mathbf{x}} \approx \mathbf{f}(\mathbf{x}^*, \mathbf{u}, t) + \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}^*} \cdot \Delta\mathbf{x} + \mathbf{w}(t) \quad (9.1.6)$$

$$\mathbf{z} \approx \mathbf{h}(\mathbf{x}^*, t) + \left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}^*} \cdot \Delta\mathbf{x} + \mathbf{v}(t) \quad (9.1.7)$$

where

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}; \quad \frac{\partial \mathbf{h}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \dots \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (9.1.8)$$

It is customary to choose the nominal trajectory $\mathbf{x}^*(t)$ to satisfy the deterministic differential equation

$$\dot{\mathbf{x}}^* = \mathbf{f}(\mathbf{x}^*, \mathbf{u}, t) \quad (9.1.9)$$

Substituting this into Eq. (9.1.6) then leads to the linearized model

$$\Delta\dot{\mathbf{x}} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}^*} \cdot \Delta\mathbf{x} + \mathbf{w}(t), \quad (\text{linearized dynamics}) \quad (9.1.10)$$

$$[\mathbf{z} - \mathbf{h}(\mathbf{x}^*, t)] = \left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}^*} \cdot \Delta\mathbf{x} + \mathbf{v}(t), \quad (\text{linearized measurement equation}) \quad (9.1.11)$$

The linearized process state equation is then

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega_0^2 & 0 & 0 & 2R_0\omega_0 \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2\omega_0}{R_0} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ w_r(t) \\ 0 \\ w_\theta(t) \end{bmatrix} \quad (9.1.23)$$

The linearized measurement equation may now be obtained either by similar incremental methods or by evaluating the partial derivative matrix indicated in Eq. (9.1.8). For variety, we choose the latter approach. From the geometry of Fig. 9.1, the two measurement relationships are seen to be

$$\begin{bmatrix} \gamma \\ \alpha \end{bmatrix} = \begin{bmatrix} \sin^{-1}\left(\frac{R_e}{r}\right) \\ \alpha_0 - \theta \end{bmatrix} \quad (9.1.24)$$

Now, treating r , \dot{r} , θ , and $\dot{\theta}$ as the total state variables leads to the partial derivative matrix

$$\begin{aligned} \left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right]_{\mathbf{x}=\mathbf{x}^*} &= \begin{bmatrix} -\frac{R_e}{\sqrt{r^2 - R_e^2}} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}_{r=R_0} \\ &= \begin{bmatrix} -\frac{R_e}{\sqrt{R_0^2 - R_e^2}} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix} \quad (9.1.25) \end{aligned}$$

This then becomes the equivalent \mathbf{H} matrix of the linearized system. The linearized model is now complete. If the measurement sequence is discrete rather than continuous, the discretization of the problem follows the same procedure as in any other linear problem: therefore, it is not necessary to pursue this further. ■

The Extended Kalman Filter

The extended Kalman filter is similar to a linearized Kalman filter except that the linearization takes place about the filter's estimated trajectory rather than a precomputed nominal trajectory. That is, the partial derivatives of Eq. (9.1.8) are evaluated along a trajectory that has been updated with the filter's estimates; these, in turn, depend on the measurements, so the filter gain sequence will depend on the sample measurement sequence realized on a particular run of the experiment. Thus, the gain sequence is not predetermined by the process model assumptions as in the usual Kalman filter.

A general analysis of the extended Kalman filter is difficult because of the feedback of the measurement sequence into the process model. However, qualitatively it would seem to make sense to update the trajectory that is used for the linearization—after all, why use the old trajectory when a better one is available? The flaw in this argument is this: the “better” trajectory is only better in a *statistical* sense. There is a chance (and maybe a good one) that the updated trajectory will be poorer than the nominal one. In that event, the estimates may be poorer; this, in turn, leads to further error in the trajectory which causes further errors in the estimates, and so forth and so forth, leading to eventual divergence of the filter. The net result is that the extended Kalman filter is a somewhat riskier filter than the regular linearized filter, especially in situations where the initial uncertainty and measurement errors are large. It may be better *on the average* than the linearized filter, but it is also more likely to diverge in unusual situations.

The aided inertial navigation example discussed in Section 6.4 may be used to illustrate the difference between the extended and regular linearized Kalman filters. In this application we may think of the inertial measurements (including their errors) as providing a nominal trajectory in state space. Strictly speaking, these inputs are not deterministic in the usual precomputable sense. But, as they unfold in real time, they become known inputs that, when appropriately processed by the system computer, yield a computed trajectory in much the same manner as would be obtained in an off-line problem. The computed trajectory, however, differs from the true one because of random instrument errors. If we use the raw uncorrected inertial outputs for the on-line computation of the filter parameters (such as those that depend on latitude and longitude), we have a simple linearized Kalman filter. This corresponds to the feed-forward configuration shown in Fig. 6.2. On the other hand, if we use the information from the aiding sources to correct the inertial system internally, and if the “corrected” inertial outputs are used to compute the filter parameters, we then have an extended Kalman filter. It should be noted in this example that the measurement for the Kalman filter is the difference between the total aiding measurement and the corresponding quantity based on the computed trajectory, just as indicated in Eq. (9.1.11).

Thus we see that the nonlinear filter viewpoint provides an alternative way of looking at the overall integrated navigation system problem. Both the extended and linearized Kalman filters have their place in this application. In cases where the mission time is short and initial uncertainties are high (or measurements are sparse), one might well choose the regular linearized filter. On the other hand, if the mission time is long (e.g., months at sea in the marine application) one almost has to use an extended Kalman filter. No general statements can be made that are foolproof in this regard, though. Each case must be considered on its own merits, and alternative viewpoints are helpful in this regard.

9.2 ADAPTIVE (SELF-LEARNING) KALMAN FILTER

In the usual Kalman filter we assume that all the process parameters, that is, ϕ_k , \mathbf{H}_k , \mathbf{R}_k , and \mathbf{Q}_k , are known. They may vary with time (index k) but, if so, the nature of the variation is assumed to be known. In physical problems this is often a rash assumption. There may be large uncertainty in some parameters because of inadequate prior test data about the process. Or, some parameter might be expected to change slowly with time, but the exact nature of the change is not predictable. In such cases, it is highly desirable to design the filter to be self-learning, so that it can adapt itself to the situation at hand, whatever that might be. This problem has received considerable attention since Kalman's original papers of the early 1960s. However, it is not an easy problem with one simple solution. This is evidenced by the fact that 20 years later we still see occasional papers on the subject in current control system journals. We will confine our attention here to a solution first presented by D. T. Magill (6). This approach leads to a somewhat unwieldy solution for many applications, but it is important because of its generality, and it serves as a departure point for other more intuitive adaptive schemes.

We begin with the simple statement that the desired estimator is to be the conditional mean given by

$$\hat{\mathbf{x}}_k = \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\mathbf{z}_k^*) d\mathbf{x} \quad (9.2.1)$$

where \mathbf{z}_k^* denotes all the measurements up to and including time t_k (i.e., \mathbf{z}_1 , \mathbf{z}_2 , ..., \mathbf{z}_k), and $p(\mathbf{x}|\mathbf{z}_k^*)$ is the probability density function of \mathbf{x}_k with the conditioning shown in parentheses.* The indicated integration is over the entire \mathbf{x} space. If all the processes in question are Gaussian, we are assured that the conditional mean will also lead to minimum mean-square error. Now, we also wish to assume that some parameter of the process, say α , is unknown to the observer. We assume that it is a random variable; thus, on any particular sample run it will be an unknown constant, but with a known statistical distribution. Thus, in reality, $p(\mathbf{x}|\mathbf{z}_k^*)$ in Eq. (9.2.1) is the joint density $p(\mathbf{x}, \alpha|\mathbf{z}_k^*)$. We are still looking for the mean of \mathbf{x}_k conditioned only on \mathbf{z}_k^* (and not α); thus, the α dependence must be summed out. Therefore, we have

$$\hat{\mathbf{x}}_k = \int_{\mathbf{x}} \mathbf{x} \int_{\alpha} p(\mathbf{x}, \alpha|\mathbf{z}_k^*) d\alpha d\mathbf{x} \quad (9.2.2)$$

* Throughout this section we will use a looser notation than that used in Chapter I in that p will be used for both probability density and discrete probability. In this way we avoid the multitudinous subscripts that would otherwise be required for conditioned multivariate random variables. However, this means that the student must use a little imagination and interpret the symbol p properly within the context of its use in any particular derivation.

But the joint density in Eq. (9.2.2) can be written as

$$p(\mathbf{x}, \alpha|\mathbf{z}_k^*) = p(\mathbf{x}|\alpha, \mathbf{z}_k^*) p(\alpha|\mathbf{z}_k^*) \quad (9.2.3)$$

Substituting Eq. (9.2.3) into (9.2.2) and interchanging the order of integration leads to

$$\hat{\mathbf{x}}_k = \int_{\alpha} p(\alpha|\mathbf{z}_k^*) \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\alpha, \mathbf{z}_k^*) d\mathbf{x} d\alpha \quad (9.2.4)$$

The inner integral will be recognized as just the usual Kalman filter estimate for a given α . This is denoted as $\hat{\mathbf{x}}_k(\alpha)$ where α shown in parentheses is intended as a reminder that there is α dependence. Equation (9.2.4) may now be rewritten as

$$\hat{\mathbf{x}}_k = \int_{\alpha} \hat{\mathbf{x}}_k(\alpha) p(\alpha|\mathbf{z}_k^*) d\alpha \quad (9.2.5)$$

Or, the discrete random variable equivalent of Eq. (9.2.5) would be

$$\hat{\mathbf{x}}_k = \sum_{i=1}^L \hat{\mathbf{x}}_k(\alpha_i) p(\alpha_i|\mathbf{z}_k^*) \quad (9.2.6)$$

where $p(\alpha_i|\mathbf{z}_k^*)$ is the discrete probability for α_i , conditioned on the measurement sequence \mathbf{z}_k^* . We will concentrate on the discrete form from this point on in our discussion.

Equation (9.2.6) simply says that the optimal estimate is a weighted sum of Kalman filter estimates with each Kalman filter operating with a separate known value of α . This is shown in Fig. 9.2. The problem now reduces to one of determining the weight factors $p(\alpha_1|\mathbf{z}_k^*)$, $p(\alpha_2|\mathbf{z}_k^*)$, etc. These, of course, change with each recursive step as the measurement process evolves in time. Presum-

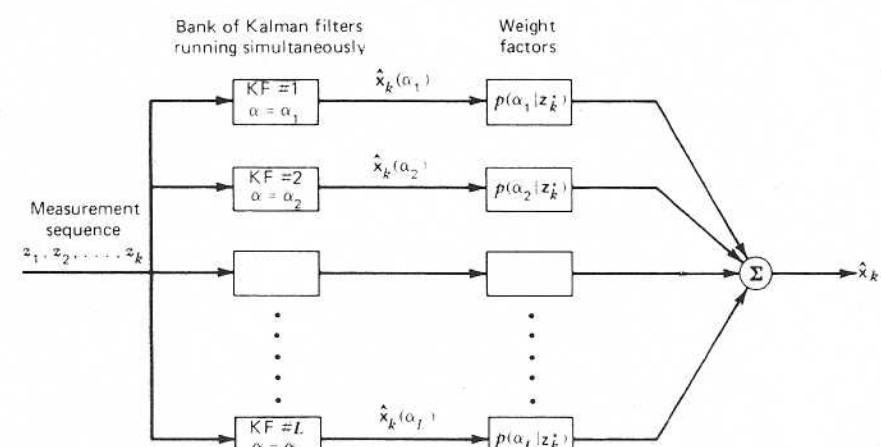


Figure 9.2. Weighted sum of Kalman filter estimates.

ably, as more and more measurements become available, we learn more about the state of the process and the unknown parameter α . (Note it is constant for any particular sample run of the process.)

We now turn to the matter of finding the weight factors indicated in Fig. 9.2. Toward this end we use Bayes rule:

$$p(\alpha_i | \mathbf{z}_k^*) = \frac{p(\mathbf{z}_k^* | \alpha_i)p(\alpha_i)}{p(\mathbf{z}_k^*)} \quad (9.2.6)$$

But,

$$\begin{aligned} p(\mathbf{z}_k^*) &= \sum_{j=1}^L p(\mathbf{z}_k^*, \alpha_j) \\ &= \sum_{j=1}^L p(\mathbf{z}_k^* | \alpha_j)p(\alpha_j) \end{aligned} \quad (9.2.7)$$

Equation (9.2.7) may now be substituted into Eq. (9.2.6) with the result

$$p(\alpha_i | \mathbf{z}_k^*) = \left[\frac{p(\mathbf{z}_k^* | \alpha_i)p(\alpha_i)}{\sum_{j=1}^L p(\mathbf{z}_k^* | \alpha_j)p(\alpha_j)} \right], \quad i = 1, 2, \dots, L \quad (9.2.8)$$

Presumably the distribution $p(\alpha_i)$ is known, and so it remains to determine $p(\mathbf{z}_k^* | \alpha_i)$ in Eq. (9.2.8). We assume all processes to be Gaussian. We further assume the system to be observable and, in the interest of simplicity, we let the measurements be scalar. We can then write $p(\mathbf{z}_k^* | \alpha_i)$ as a multivariate normal density function of the form

$$p(\mathbf{z}_k^* | \alpha_i) = \frac{1}{(2\pi)^{k/2} |\mathbf{C}(i)|^{1/2}} \exp \left\{ -\frac{1}{2} [\mathbf{z}_k^{*T} \mathbf{C}(i)^{-1} \mathbf{z}_k^*] \right\}, \quad i = 1, 2, \dots, L \quad (9.2.9)$$

where $\mathbf{C}(i)$ is the covariance matrix associated with $\mathbf{z}_k, \mathbf{z}_{k-1}, \dots, \mathbf{z}_1$ conditioned on α_i .

We now need to find $\mathbf{C}(i)$ in terms of quantities normally computed on-line in the bank of Kalman filters. To do this we first write $p(\mathbf{z}_k^* | \alpha_i)$ as a product of conditional density functions. Temporarily omitting the α_i conditioning (just to save writing), we have

$$\begin{aligned} p(\mathbf{z}_k^*) &= p(\mathbf{z}_k, \mathbf{z}_{k-1}, \dots, \mathbf{z}_1) \\ &= p(\mathbf{z}_k, \mathbf{z}_{k-1}, \dots, \mathbf{z}_2 | \mathbf{z}_1)p(\mathbf{z}_1) \\ &= p(\mathbf{z}_k, \mathbf{z}_{k-1}, \dots, \mathbf{z}_3 | \mathbf{z}_2, \mathbf{z}_1)p(\mathbf{z}_2 | \mathbf{z}_1)p(\mathbf{z}_1) \\ &\quad \cdot \\ &\quad \cdot \\ &= p(\mathbf{z}_k | \mathbf{z}_{k-1}, \mathbf{z}_{k-2}, \dots, \mathbf{z}_1)p(\mathbf{z}_{k-1} | \mathbf{z}_{k-2}, \mathbf{z}_{k-3}, \dots, \mathbf{z}_1) \dots p(\mathbf{z}_2 | \mathbf{z}_1)p(\mathbf{z}_1) \end{aligned} \quad (9.2.10)$$

Note that each factor in Eq. (9.2.10) is a probability density of a measurement, given all the previous measurements. Each of these must be normal in form. In addition, the measurements \mathbf{z}_j are linearly related to the process by the measurement equation $\mathbf{z}_j = \mathbf{H}_j \mathbf{x}_j + v_j$. Thus the covariance matrix associated with the product form given by Eq. (9.2.10) is

$$\mathbf{C}'(i) = \begin{bmatrix} (\mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k^T + R_k)_i & \mathbf{0} \\ (\mathbf{H}_{k-1} \mathbf{P}_{k-1}^- \mathbf{H}_{k-1}^T + R_{k-1})_i & \cdot \\ \vdots & \vdots \\ \mathbf{0} & (\mathbf{H}_1 \mathbf{P}_1^- \mathbf{H}_1^T + R_1)_i \end{bmatrix} \quad (9.2.11)$$

The super minus indicates "a priori" just as in previous chapters. The terms along the diagonal of Eq. (9.2.11) will be recognized as quantities used in the gain computation in the normal Kalman filter. Thus the computation of $|\mathbf{C}'(i)|$ is relatively simple. Each elemental filter in the bank of Kalman filters keeps updating its own running product of terms of the form $(\mathbf{H}\mathbf{P}^- \mathbf{H}^T + R)$ with each step of the process. These running products will, of course, be different for each elemental filter because of the different values of the parameter α used in each filter. (Note that the $(\mathbf{H}\mathbf{P}^- \mathbf{H}^T + R)$ terms are not a function of the measurement data; therefore they could be precomputed, if so desired.)

We now turn to the $\mathbf{z}_k^{*T} \mathbf{C}(i)^{-1} \mathbf{z}_k^*$ term that appears in the exponent in Eq. (9.2.9). Since the $\mathbf{C}'(i)$ matrix is diagonal, its inverse is a diagonal matrix with scalar reciprocals along the diagonal. Recall that the variates that appear in the probability densities of Eq. (9.2.10) are the k measurements, each of which is conditioned on all previous measurements. Thus each variate is a biased Gaussian random variable, with the bias being simply the a priori estimate of the random variable. Thus we have (with the i index omitted for compactness)

$$\begin{aligned} \mathbf{z}_k^{*T} \mathbf{C}^{-1} \mathbf{z}_k^* &= \frac{(\mathbf{z}_k - \hat{\mathbf{z}}_k^-)^2}{\mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k^T + R_k} + \frac{(\mathbf{z}_{k-1} - \hat{\mathbf{z}}_{k-1}^-)^2}{\mathbf{H}_{k-1} \mathbf{P}_{k-1}^- \mathbf{H}_{k-1}^T + R_{k-1}} \\ &\quad + \dots + \frac{(\mathbf{z}_1 - \hat{\mathbf{z}}_1^-)^2}{\mathbf{H}_1 \mathbf{P}_1^- \mathbf{H}_1^T + R_1} \end{aligned} \quad (9.2.12)$$

The exponent term $\mathbf{z}_k^{*T} \mathbf{C}^{-1} \mathbf{z}_k^*$ can now be seen to be a weighted sum of the squared measurement residuals. Again, each filter in the bank of Kalman filters must keep updating its own sum-of-residuals quantity with each step of the process. This is relatively easy, though, because the Kalman filter normally computes the measurement residual and the $(\mathbf{H}\mathbf{P}^- \mathbf{H}^T + R)$ quantity in the course of its normal operation.

Finally, referring to Eq. (9.2.8), we see that the filter weight factors are formed as follows:

1. Each filter computes its own $p(\mathbf{z}_k^*|\alpha_i)$ from Eq. (9.2.9). This involves keeping a running product of $(\mathbf{H}\mathbf{P}^{-1}\mathbf{H}^T + R)$ terms and a running sum of weighted squared measurement residuals. These are then used in the simple exponential expression indicated in Eq. (9.2.9) with the final result being $p(\mathbf{z}_k^*|\alpha_i)$.
2. The $p(\mathbf{z}_k^*|\alpha_i)$ from step 1 are each multiplied by the known $p(\alpha_i)$ within each filter yielding $p(\mathbf{z}_k^*|\alpha_i)p(\alpha_i)$. Each filter then transmits this product to a summer, which forms

$$\sum_{j=1}^L p(\mathbf{z}_k^*|\alpha_i)p(\alpha_i)$$

3. The final weight factors for each filter are then formed by dividing the $p(\mathbf{z}_k^*|\alpha_i)p(\alpha_i)$ by the sum from step 2.

It can be seen then that computation of the conditional densities (which appeared quite formidable at first glance) works out to be relatively simple. This is because certain key quantities can be formed by updating running sums and products.

We are now in a position to reflect on the whole adaptive filter system in perspective. Qualitatively, the adaptation proceeds as follows. Prior to receiving any measurements, the system must set the weight factors equal to the a priori probabilities $p(\alpha_i)$. It has no better information about α as of this point in time. Then, as measurements are accumulated, each elemental Kalman filter sums its weighted squared residuals and uses this sum as the negative exponent in its Gaussian density computation. As time proceeds, the correct filter's residuals work out to be smaller (on the average) than the others, and thus its probability density is the largest and it is given the most weight in the blending of the elemental estimates. The measurement residuals are summed, with the effect being cumulative; and, in the limit, a weight factor of unity for the correct filter (and zero for the others) is approached. In effect, the system "learns" which is the correct α , and then assigns all of the weight to this filter's estimate.

The adaptive scheme due to Magill is not without some practical problems, but it is still important because it is optimum (within the Gaussian assumption), and it serves as a point of departure for other less rigorous approaches. For example, the Magill approach tells us that the measurement residuals are key test statistics to use in the adaptation process. In this regard, the derivation serves as an indirect justification of the principle that an optimal filter minimizes the expectation of the squared residuals (see Problem 9.3). With this in mind, one can think of many less-complicated ways of using the measurement residuals to make the filter adapt to the correct α . After all, the system only has to implement the various options and choose the one with the smallest average residuals. Simpler intuitive algorithms might not converge quite as rapidly as the optimal scheme, but they might well be considerably easier to implement.

In the foregoing discussion, remember that we assumed α to be constant

with time. Thus we should not expect this adaptive filter to adjust itself to a change in α after it has been in operation for a long time. In its ideal form, the filter never forgets—no matter how old the measurement data! If a time-variable adaptation capability is desired, the filter algorithm obviously must be modified to limit its memory of past residuals. This might also be desirable for computational reasons. In any event, such modified forms of the basic Magill scheme are another story, and this is where we leave the subject of adaptive filtering. Also, many other adaptive schemes have been proposed, and one has only to browse through recent control system literature to appreciate the breadth and variety of schemes that have been suggested.

9.3 SQUARE ROOT FILTERING

Computational problems associated with discrete Kalman filtering were mentioned briefly in Chapter 6. Of the various difficulties, perhaps propagation of the \mathbf{P} matrix gives most trouble. If the roundoff error here is cumulative with each recursive step, then the gain computation suffers and eventually all sorts of disastrous things can happen to the filter. In off-line analysis the analyst can usually guard against this by using double- or even triple-precision arithmetic. However, this luxury is usually not possible in on-line applications.

The basic problem in propagating \mathbf{P} lies in the wide range of numerical values that the computer must accommodate to in some situations. For example, in an inertial platform alignment situation, the initial uncertainty (after coarse alignment) might be of the order of a few degrees. The final precision alignment after Kalman filtering the measurement data might be of the order of a fraction of an arc second. This would be roughly four orders of magnitude change from beginning to end in terms of standard deviation; however, in terms of variance, it would be a change of eight orders of magnitude! Accommodation to this wide dynamic range with fixed-point arithmetic (which is often dictated in on-line applications) is almost impossible in most situations. Obviously, it would be better to propagate standard deviation rather than variance in this situation. Thus various algorithms have been worked out for propagating something roughly equivalent to the matrix square root of \mathbf{P} rather than \mathbf{P} itself. These are usually lumped together into a category that now is generally referred to as square root filtering. Recent books by Bierman (7) and Maybeck (8) give good accounts of the development of square root filtering, which apparently dates back to a 1964 paper by J. E. Potter (9). We are content here to develop only the simplest case, that where the measurements are scalar and the process is undriven, that is, $\mathbf{Q}_k = 0$.

Before developing the square root filter algorithm, we digress for a moment and elaborate on the concept of a matrix square root. Let \mathbf{A} be a symmetric, positive semidefinite matrix, and assume for the moment that it can be

written in factored form as

$$\mathbf{A} = \mathbf{B}\mathbf{B}^T \quad (9.3.1)$$

\mathbf{B} may then be thought of as the “square root” of \mathbf{A} . However, Eq. (9.3.1) does not, in itself, uniquely define \mathbf{B} ; nor does it tell us how to compute the elements of \mathbf{B} , given those of \mathbf{A} . Of the many forms of \mathbf{B} that will satisfy Eq. (9.3.1), we pay particular attention to a lower triangular form that is obtained from the Cholesky decomposition algorithm. The lower triangular form for \mathbf{B} is unique and its elements may be readily obtained from those of \mathbf{A} .

The Cholesky decomposition algorithm is perhaps best described by means of a simple 3×3 example. Let Eq. (9.3.1) be written in component form as follows (for a 3×3 symmetric \mathbf{A} matrix):

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{bmatrix} b_{11} & b_{21} & b_{31} \\ 0 & b_{22} & b_{32} \\ 0 & 0 & b_{33} \end{bmatrix} \quad (9.3.2)$$

Clearly, after multiplying \mathbf{B} and \mathbf{B}^T and equating corresponding terms in the top row, we get

$$a_{11} = b_{11}^2 \quad (9.3.3)$$

$$a_{12} = b_{11}b_{21} \quad (9.3.4)$$

$$a_{13} = b_{11}b_{31} \quad (9.3.5)$$

The b_{11} term is evaluated first as $\sqrt{a_{11}}$. Then b_{21} and b_{31} may be determined as a_{12}/b_{11} and a_{13}/b_{11} . Next, equating corresponding terms in the second row in Eq. (9.3.2) yields

$$a_{22} = b_{21}^2 + b_{31}^2 \quad (9.3.6)$$

$$a_{23} = b_{21}b_{31} + b_{22}b_{32} \quad (9.3.7)$$

Having already determined b_{21} and b_{31} , clearly b_{22} and b_{32} may be determined from Eqs. (9.3.6) and (9.3.7). Finally, b_{33} is determined by equating the 33 terms on both sides of Eq. (9.3.2). Note that no simultaneous equations need be solved if the elements of \mathbf{B} are determined in the order just described. This is the beauty of the Cholesky decomposition. Also note that a scalar square root operation is necessary in a number of places in the decomposition procedure. Nonnegative real roots may always be chosen because of the positive semi-definite requirement on \mathbf{A} . Thus the diagonal elements of \mathbf{B} will be non-negative.

Returning now to the square root filter problem, we assume that the initial a priori \mathbf{P} matrix is written in factored form. To be specific, a lower triangular Cholesky decomposition will be assumed. Thus we write \mathbf{P}_0^- in the form

$$\mathbf{P}_0^- = \mathbf{W}_0^-\mathbf{W}_0^{-T} \quad (9.3.8)$$

We consider the gain equation first, and we temporarily omit subscripts just to save writing. The Kalman gain may be written as

$$\begin{aligned} \mathbf{K} &= \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + R)^{-1} \\ &= \mathbf{W}^-\mathbf{W}^{-T}\mathbf{H}^T(\mathbf{H}\mathbf{W}^-\mathbf{W}^{-T}\mathbf{H}^T + R)^{-1} \\ &= \mathbf{W}^-(\mathbf{H}\mathbf{W}^-)^T/[(\mathbf{H}\mathbf{W}^-)(\mathbf{H}\mathbf{W}^-)^T + R] \end{aligned} \quad (9.3.9)$$

Note that \mathbf{P}^- does not appear in the gain equation, and that the indicated division in Eq. (9.3.9) is possible because the measurement is scalar.

We now turn to the error covariance update equation, and the form given by Eq. (5.4.24) is repeated here for convenience (without subscripts).

$$\mathbf{P} = \mathbf{P}^- - \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + R)^{-1}\mathbf{H}\mathbf{P}^- \quad (9.3.10)$$

Since the bracketed quantity and $(\mathbf{H}\mathbf{W}^-)$ in Eq. (9.3.9) appear frequently in the following derivation, they are assigned new symbols.

$$\mathbf{H}\mathbf{W}^- = \mathbf{S} \quad (\text{a } 1 \times n \text{ matrix}) \quad (9.3.11)$$

$$[(\mathbf{H}\mathbf{W}^-)(\mathbf{H}\mathbf{W}^-)^T + R] = r \quad (\text{a scalar}) \quad (9.3.12)$$

Recalling that the factored forms of \mathbf{P} and \mathbf{P}^- are $\mathbf{W}\mathbf{W}^T$ and $\mathbf{W}^-\mathbf{W}^{-T}$, Eq. (9.3.10) may be written as

$$\begin{aligned} \mathbf{W}\mathbf{W}^T &= \mathbf{W}^-\mathbf{W}^{-T} - \mathbf{W}^-\mathbf{W}^{-T}\mathbf{H}^T(\mathbf{H}\mathbf{W}^-\mathbf{W}^{-T}\mathbf{H}^T + R)^{-1}\mathbf{H}\mathbf{W}^-\mathbf{W}^{-T} \\ &= \mathbf{W}^-\mathbf{W}^{-T} - \mathbf{W}^-\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + R)^{-1}\mathbf{S}\mathbf{W}^{-T} \\ &= \mathbf{W}^-(\mathbf{I} - \mathbf{S}^T\mathbf{S}/r)\mathbf{W}^{-T} \end{aligned} \quad (9.3.13)$$

The problem now is to write Eq. (9.3.13) in factored form. In particular, the inner quantity in parentheses must be factored. We will try the form

$$(\mathbf{I} - \mathbf{S}^T\mathbf{S}/r) = (\mathbf{I} - a\mathbf{S}^T\mathbf{S})(\mathbf{I} - a\mathbf{S}\mathbf{S}^T)^T \quad (9.3.14)$$

where a is a scalar yet to be determined. Expanding the right side of Eq. (9.3.14) yields

$$(\mathbf{I} - \mathbf{S}^T\mathbf{S}/r) = \mathbf{I} - a\mathbf{S}^T\mathbf{S} - a\mathbf{S}^T\mathbf{S} + a^2\mathbf{S}^T\mathbf{S}\mathbf{S}^T \quad (9.3.15)$$

We now note that $\mathbf{S}\mathbf{S}^T$ is scalar, and thus Eq. (9.3.15) may be rewritten in the form

$$\left(a^2\mathbf{S}\mathbf{S}^T - 2a + \frac{1}{r}\right)\mathbf{S}^T\mathbf{S} = 0 \quad (9.3.16)$$

Now, since $\mathbf{S}^T\mathbf{S}$ is not zero, for equality to hold in Eq. (9.3.16), we must have

$$a^2\mathbf{S}\mathbf{S}^T - 2a + \frac{1}{r} = 0 \quad (9.3.17)$$

This quadratic equation in a may now be solved. The result is

$$a = \frac{1 \pm \sqrt{1 - \mathbf{S}\mathbf{S}^T/r}}{\mathbf{S}\mathbf{S}^T} \quad (9.3.18)$$

We now note that $r = \mathbf{S}\mathbf{S}^T + R$. Thus, our solution for a may be written in the form

$$a = \frac{1}{\mathbf{S}\mathbf{S}^T} \left(1 \pm \sqrt{\frac{R}{r}} \right) \quad (9.3.19)$$

Finally, we choose the negative sign in Eq. (9.3.19) in order to avoid large a as $\mathbf{S}\mathbf{S}^T$ becomes small. Thus our final solution for a is

$$a = \frac{1}{\mathbf{S}\mathbf{S}^T} \left(1 - \sqrt{\frac{R}{r}} \right) \quad (9.3.20)$$

We refer back to Eqs. (9.3.13) and (9.3.14) to obtain the factored form for \mathbf{P} .

$$\mathbf{W}\mathbf{W}^T = \mathbf{W}^-(\mathbf{I} - a\mathbf{S}^T\mathbf{S})(\mathbf{I} - a\mathbf{S}^T\mathbf{S})^T \mathbf{W}^{-T} \quad (9.3.21)$$

The update equation for \mathbf{W} is then

$$\mathbf{W} = \mathbf{W}^-(\mathbf{I} - a\mathbf{S}^T\mathbf{S}) \quad (9.3.22)$$

where

$$\mathbf{S} = \mathbf{H}\mathbf{W}^- \quad (9.3.23)$$

and a is given by Eq. (9.3.20).

The projection step of the square root algorithm is relatively simple. Recall that the process was assumed to be undriven; thus $\mathbf{Q}_k = 0$. \mathbf{P}_{k+1}^- is then given by

$$\mathbf{P}_{k+1}^- = \boldsymbol{\phi}_k \mathbf{P}_k \boldsymbol{\phi}_k^T \quad (9.3.24)$$

Or, in terms of \mathbf{W} we have

$$\begin{aligned} \mathbf{W}_{k+1}^- \mathbf{W}_{k+1}^{-T} &= \boldsymbol{\phi}_k \mathbf{W}_k \mathbf{W}_k^T \boldsymbol{\phi}_k^T \\ &= (\boldsymbol{\phi}_k \mathbf{W}_k)(\boldsymbol{\phi}_k \mathbf{W}_k)^T \end{aligned} \quad (9.3.25)$$

Clearly, the appropriate projection equation is

$$\mathbf{W}_{k+1}^- = \boldsymbol{\phi}_k \mathbf{W}_k \quad (9.3.26)$$

The square root recursive algorithm is now complete, and it is summarized (with subscripts reinserted) in Fig. 9.3.

A few comments about the square root algorithm are in order. First, even though this is a restricted case, it is of considerable practical importance. The scalar measurement situation is quite common, and the undriven process case (e.g., random constants) is exactly the one that presents problems in terms of extreme dynamic range of the \mathbf{P} matrix. It will suffice to say that more general

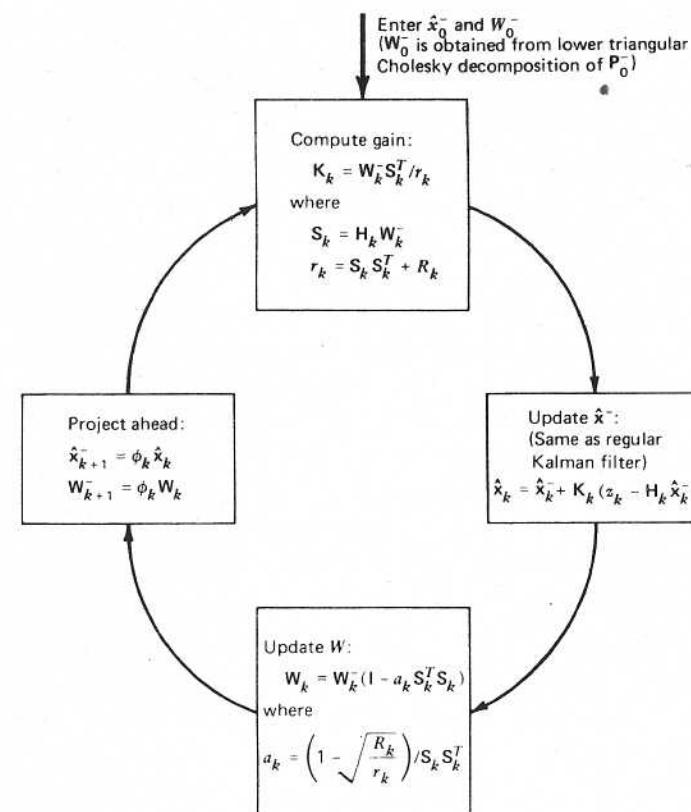


Figure 9.3. Square root Kalman filter algorithm for the special case for $\mathbf{Q}_k = 0$ and measurements are scalar.

square root algorithms have been developed, and Bierman (7) gives a good account of these.

Elaborating further on the shortcomings of our basic algorithm, we note that even though we begin the algorithm with a lower triangular \mathbf{W}_0^- matrix (for convenience in factoring \mathbf{P}_0^-), we quickly lose the triangular form in the \mathbf{W} -update and projection steps. Thus \mathbf{W} is generally a full matrix with no obvious symmetry, and n^2 elements must be propagated. This is in contrast to the \mathbf{P} -matrix form which has symmetry and has only $n(n + 1)/2$ independent elements. In addition, a scalar square root operation must be performed with each recursive cycle. This is a time-consuming arithmetic operation, and thus there is some disadvantage in this regard.

A newer factorization technique known as **U-D** factorization avoids the square root problem, and thus it has received considerable attention in recent

years. Basically, **U-D** factorization involves factoring the **P** matrix into the form

$$\mathbf{P} = \mathbf{UDU}^T$$

where **D** is diagonal and **U** is unit upper triangular (**I**'s along the diagonal). The nontrivial elements of **U** and **D** are then propagated. This algorithm is reported to be superior in most respects to the basic square root algorithm presented here. It is more complicated to develop, though, and will not be pursued further here. Either Bierman (7) or Maybeck (8) may be consulted for the details of **U-D** factorization.

The point of all this discussion is simply this. If propagation of the **P** matrix gives trouble in a particular application, one does not have to be content with the usual Kalman filter recursive equations. Various square root algorithms have been developed that have better numerical characteristics. However, since they are somewhat more complicated, in off-line analysis one usually uses the regular algorithm with high-precision arithmetic (if necessary). On the other hand, in on-line applications it might be another matter. Use of a square root algorithm might well make the difference between a viable estimator and one that is not.

9.4. CONVERSION OF CONTINUOUS MEASUREMENTS TO DISCRETE FORM—THE DELAYED STATE FILTER ALGORITHM

The continuous Kalman filter discussed in Chapter 7 is of considerable theoretical importance, but it would be a rare situation where one would actually implement the filter in analog form. Usually, in the end, the analyst/designer must make the transition to the discrete form, even though some (or perhaps all) of the measurements are continuous. Fortunately the discrete filter equations are exact in their own right and are not approximations to the continuous case. Thus there is no error in making the transition from the continuous to the discrete case relative to the recursive equations. However, in the transition the continuous measurements must be converted to equivalent sampled data (in some sense of equivalence), and there are subtleties associated with this conversion that warrant further comment.

To illustrate the problems that can arise in the continuous-to-discrete conversion, consider a scalar Gauss-Markov process whose spectral content is of the order of a few hertz. Let the additive noise be Gaussian and flat out to a few kilohertz. In order not to lose any signal information, it is tempting to sample

the measurement as fast as possible, perhaps in the kilohertz range. However, this might pose unreasonable demands on the system computer. Certainly, we do not want to force the filter to cycle through millions of recursive steps when thousands would do just as well! On the other extreme, if we sample too slowly, say at less than the signal Nyquist rate, we should expect to lose signal information; and, as a result, the filter's performance will be poorer than need be. Thus, the sampling should be treated with some care. One technique that has been used successfully is to sample the measurement at a fairly high rate (if the frequency content of the noise so dictates), and then preaverage the samples over the Kalman-filter update interval to form a discrete measurement for the filter. In the example at hand, it might be reasonable to sample the continuous measurement at a rate of 1000 Hz and then preaverage 100-sample blocks to form "compressed" samples to present to the Kalman filter every .1 sec. The error variance associated with the compressed samples would then be less than that of the raw samples by a factor of 100, provided the raw sample errors were independent. Of course, the R_k parameter of the Kalman filter would be the variance associated with the compressed samples rather than that of the raw samples.

The preaveraging technique just described works fine in many applications and nothing more sophisticated is required. However, there are cases where, for one reason or another, the Kalman filter must be relatively slow, and the resulting prefiltering interval is longer than desired. This presents problems for two reasons. First, some measurement information will be lost due to the low sampling rate. The loss may not be serious, though, if the sampling rate is not too low relative to the signal frequency content. A more serious problem, though, may lie in the modeling inaccuracy that comes about from the assumption that the signal is constant over the averaging interval; when, in fact, it is not. In averaging, we integrate (or sum in the discrete case) over the update interval, and hence the compressed measurement presented to the filter is, in reality, of the form

$$\mathbf{z}_k = \int_{t_{k-1}}^{t_k} [\mathbf{Hx}(t) + \mathbf{v}(t)] dt \quad (9.4.1)$$

If $\mathbf{x}(t)$ varies appreciably over the interval, the resultant \mathbf{z}_k may be more representative of the value of $\mathbf{x}(t)$ at the midpoint than at the endpoint t_k , which is, of course, the assumption in the Kalman filter model. When this situation arises, there is a method of modifying the recursive equations to accommodate exactly to measurements of the form given by Eq. (9.4.1). This leads to the concept of the delayed-state filter algorithm, which we introduce by means of a simple example.

Assume for the moment that position and velocity are two of the state variables of the system. Suppose further that we have a continuous measurement of velocity that we wish to convert to discrete form (not an uncommon problem). Assume that we integrate (average) the noisy velocity measurement

over the update interval. We then have

$$\begin{aligned} z_k &= \int_{t_{k-1}}^t (\text{Velocity} + \text{noise}) dt = (\text{Position at } t_k) \\ &\quad - (\text{Position at } t_{k-1}) + (\text{Residual error}) \end{aligned} \quad (9.4.2)$$

Thus we find that z_k has a linear connection to both the present and previous states, rather than just the present state as assumed in the usual Kalman filter model. In more general terms, the measurement situation in this case is of the form

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{J}_k \mathbf{x}_{k-1} + \mathbf{v}_k \quad (9.4.3)$$

We now seek an appropriate set of recursive equations for the measurement model given by Eq. (9.4.3). As before, we assume \mathbf{v}_k to be a white sequence with a covariance matrix \mathbf{R}_k . The desired solution is worked out in references 10 and 11 by combining the present and previous state vectors into a composite state vector with twice the dimension of the original ones. Here, we can accomplish the same result by a more direct path by taking advantage of the material in Chapter 5.

We begin with the estimate update equation

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k-} + \mathbf{K}_k(z_k - \mathbf{H}_k \hat{\mathbf{x}}_{k-} - \mathbf{J}_k \hat{\mathbf{x}}_{k-1}) \quad (9.4.4)$$

Note that our best estimate of \mathbf{z}_k prior to receiving the measurement is $(\mathbf{H}_k \hat{\mathbf{x}}_{k-} + \mathbf{J}_k \hat{\mathbf{x}}_{k-1})$, rather than just $\mathbf{H}_k \hat{\mathbf{x}}_{k-}$ as in the usual Kalman filter. Next, we form the error and error covariance expressions just as was done in Chapter 5 [see Eqs. (5.4.9) to (5.4.11)].

$$\begin{aligned} \mathbf{e}_k &= \mathbf{x}_k - \hat{\mathbf{x}}_k \\ &= \mathbf{e}_{k-} - \mathbf{K}_k(\mathbf{H}_k \mathbf{e}_{k-} + \mathbf{J}_k \mathbf{e}_{k-1} + \mathbf{v}_k) \end{aligned} \quad (9.4.5)$$

And

$$\begin{aligned} \mathbf{P}_k &= E(\mathbf{e}_k \mathbf{e}_k^T) \\ &= E[(\mathbf{e}_{k-} - \mathbf{K}_k(\mathbf{H}_k \mathbf{e}_{k-} + \mathbf{J}_k \mathbf{e}_{k-1} + \mathbf{v}_k)) \\ &\quad [\mathbf{e}_{k-} - \mathbf{K}_k(\mathbf{H}_k \mathbf{e}_{k-} + \mathbf{J}_k \mathbf{e}_{k-1} + \mathbf{v}_k)]^T] \end{aligned} \quad (9.4.6)$$

We now note that

$$\begin{aligned} \mathbf{e}_{k-} &= \mathbf{x}_k - \hat{\mathbf{x}}_{k-} \\ &= \boldsymbol{\phi}_{k-1} \mathbf{x}_{k-1} + \mathbf{w}_{k-1} - \boldsymbol{\phi}_{k-1} \hat{\mathbf{x}}_{k-1} \\ &= \boldsymbol{\phi}_{k-1} \mathbf{e}_{k-1} + \mathbf{w}_{k-1} \end{aligned} \quad (9.4.7)$$

We can also take advantage of the knowledge that \mathbf{v}_k is uncorrelated with \mathbf{e}_k and \mathbf{e}_{k-1} , and \mathbf{w}_{k-1} is uncorrelated with \mathbf{e}_{k-1} . Then, expanding Eq. (9.4.6) and

performing the indicated expectations leads to (after a modest amount of algebra)

$$\begin{aligned} \mathbf{P}_k &= \mathbf{P}_{k-} - \mathbf{K}_k(\mathbf{H}_k \mathbf{P}_{k-} + \mathbf{J}_k \mathbf{P}_{k-1} \boldsymbol{\phi}_{k-1}^T) - (\mathbf{P}_{k-} \mathbf{H}_k^T + \boldsymbol{\phi}_{k-1} \mathbf{P}_{k-1} \mathbf{J}_k^T) \mathbf{K}_k^T \\ &\quad + \mathbf{K}_k(\mathbf{H}_k \mathbf{P}_{k-} \mathbf{H}_k^T + \mathbf{R}_k + \mathbf{J}_k \mathbf{P}_{k-1} \mathbf{J}_k^T + \mathbf{H}_k \boldsymbol{\phi}_{k-1} \mathbf{P}_{k-1} \mathbf{J}_k^T \\ &\quad + \mathbf{J}_k \mathbf{P}_{k-1} \boldsymbol{\phi}_{k-1}^T \mathbf{H}_k^T) \mathbf{K}_k^T \end{aligned} \quad (9.4.8)$$

Equation (9.4.8) may now be compared with Eq. (5.4.13). Clearly, the forms are identical with the following replacements:

$$\mathbf{H}\mathbf{P}^- \rightarrow \mathbf{H}_k \mathbf{P}_{k-} + \mathbf{J}_k \mathbf{P}_{k-1} \boldsymbol{\phi}_{k-1}^T \quad (9.4.9)$$

$$\begin{aligned} (\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}) &\rightarrow \mathbf{H}_k \mathbf{P}_{k-} \mathbf{H}_k^T + \mathbf{R}_k + \mathbf{J}_k \mathbf{P}_{k-1} \mathbf{J}_k^T + \mathbf{H}_k \boldsymbol{\phi}_{k-1} \mathbf{P}_{k-1} \mathbf{J}_k^T \\ &\quad + \mathbf{J}_k \mathbf{P}_{k-1} \boldsymbol{\phi}_{k-1}^T \mathbf{H}_k^T \end{aligned} \quad (9.4.10)$$

The same optimization arguments used in Chapter 5 may be used here, and hence the gain expression for the delayed-state algorithm is obtained simply by making the appropriate replacements in the usual Kalman gain formula. The result is

$$\begin{aligned} \mathbf{K}_k &= (\mathbf{P}_{k-} \mathbf{H}_k^T + \boldsymbol{\phi}_{k-1} \mathbf{P}_{k-1} \mathbf{J}_k^T) [\mathbf{H}_k \mathbf{P}_{k-} \mathbf{H}_k^T + \mathbf{R}_k + \mathbf{J}_k \mathbf{P}_{k-1} \mathbf{J}_k^T \\ &\quad + \mathbf{H}_k \boldsymbol{\phi}_{k-1} \mathbf{P}_{k-1} \mathbf{J}_k^T + \mathbf{J}_k \mathbf{P}_{k-1} \boldsymbol{\phi}_{k-1}^T \mathbf{H}_k^T]^{-1} \end{aligned} \quad (9.4.11)$$

The error covariance update expression may be written in a number of different forms, just as in the usual Kalman filter. Perhaps the simplest form for the delay-state algorithm is

$$\mathbf{P}_k = \mathbf{P}_{k-} - \mathbf{K}_k \mathbf{L}_k \mathbf{K}_k^T \quad (9.4.12)$$

where \mathbf{L}_k is the bracketed term of Eq. (9.4.11) (not inverted). Since both \mathbf{K}_k and \mathbf{L}_k are computed in the gain computation step, Eq. (9.4.12) is easy to evaluate. The projection equations are the same as for the usual filter and need not be repeated.

Examples illustrating the use of the delayed state algorithm in aided inertial navigation systems are given in references 10, 11, and 12. In these examples position and velocity were "natural" system state variables. Thus measurements that were related to the integral of velocity fit the format of Eq. (9.4.3) directly without any modification of the process model. This is not always the case, though, as will be seen presently. The referenced navigation examples are interesting practical applications, but they are somewhat lengthy; therefore, we will look at a simpler example here. We consider a scalar Markov process with preaveraged samples as the measurement sequence.

Example 9.2 Consider a scalar Gauss-Markov process $x(t)$ with an auto-correlation function

$$R_x(\tau) = e^{-\tau} \quad (9.4.13)$$

From previous work in Chapter 5 (Example 5.2), we know that this process satisfies a first-order differential equation with white noise as the driving func-

tion. Suppose now that we have a continuous noisy measurement of $x(t)$ with the additive noise being Gaussian and white with a spectral amplitude of .02. The continuous Kalman filter model is then (see Chapter 7)

$$\dot{x} = -x + \sqrt{2} w(t) \quad (Q = 1) \quad (9.4.14)$$

$$z = x + v(t) \quad (R = .02) \quad (9.4.15)$$

Suppose further that we wish to implement the filter in discrete form (rather than continuous), and that the filter recursive cycle must be relatively slow in order not to impose undue hardship on the system computer (along with its other work load). Since the process has a time constant of unity, it would seem reasonable to average the measurement data over intervals of one-tenth of the time constant and then present the discrete filter with a sequence of preaveraged data at update intervals of .1 sec. In this case we know there will be some loss in measurement information due to the preaveraging; the question is, how much loss? For comparison purposes we examine three possibilities:

1. The continuous Kalman filter. This baseline system represents the best we could hope to do with an infinite sampling and update rate.
2. Optimal operation on the preaveraged data using the delayed-state Kalman filter algorithm. (There will be some degradation here as compared with the continuous case, but at least the sequence of compressed measurements will be processed optimally.)
3. The usual discrete Kalman filter which processes the preaveraged data just as if they were truly instantaneous samples of $x(t)$ at the update times t_1, t_2, \dots, t_k . (Here we have to contend with a faulty measurement model as well as preaveraged data.)

The Continuous Kalman Filter

This example is similar to the one worked out in Example 7.1. Since the R parameter is the only element that is different in these two examples, the pair of linear equations to be solved here are

$$\dot{X} = -X + 2Z, \quad X(0) = 1 \quad (9.4.16)$$

$$\dot{Z} = 50X + Z, \quad Z(0) = 1 \quad (9.4.17)$$

These equations may be solved using Laplace transform methods. The result is

$$P(t) = \frac{X(t)}{Z(t)} \approx \frac{.450248e^{-\sqrt{10}t} + .549752e^{\sqrt{10}t}}{-2.037345e^{-\sqrt{10}t} + 3.037345e^{\sqrt{10}t}} \quad (9.4.18)$$

In our comparison with the discrete models, we will be particularly interested in the error variance after a single step of .1 sec. Evaluation of Eq. (9.4.18) at $t = .1$ yields

$$P(.1) \approx .220696 \quad (9.4.19)$$

Delayed-State Kalman Filter Model

The compressed measurement in this case is related to the integral of $x(t)$ because of the averaging over the Δt interval:

$$z_k = \frac{1}{\Delta t} \int_{t_{k-1}}^{t_k} [x(t) + v(t)] dt, \quad \Delta t = .1 \text{ and} \\ k = 1, 2, \dots, n \quad (9.4.20)$$

Since this does not fit the required format of the usual discrete Kalman filter, we modify the state model to make it amenable to solution with the delayed-state algorithm. This is done by defining the integral of $x(t)$ to be a new state variable. The augmented system will then have two state variables. They are defined as follows:

$$x_1(t) = \int_0^t x_2(\tau) d\tau, \quad (\text{a new state variable}) \quad (9.4.21)$$

$$x_2(t) = x(t), \quad (\text{the original scalar Markov process}) \quad (9.4.22)$$

The details of the discrete model for this two-state process are given in Example 5.3. For a step size of .1 sec., the key parameters are

$$\phi_k = \begin{bmatrix} 1 & .0951626 \\ 0 & .9048374 \end{bmatrix} \quad (9.4.23)$$

$$Q_k = \begin{bmatrix} .00061892 & .0090559 \\ .0090559 & .1812692 \end{bmatrix} \quad (9.4.24)$$

The discrete measurement model is obtained by combining Eqs. (9.4.20), (9.4.21), and (9.4.22) for a step size of .1 sec. This yields

$$z_k = 10 \int_{t_{k-1}}^{t_k} x_2(\tau) d\tau \\ = 10 x_1(t_k) - 10 x_1(t_{k-1}) + 10 \int_{t_{k-1}}^{t_k} v(t) dt \quad (9.4.25)$$

Clearly, this is of the form given by Eq. (9.4.3) where

$$H_k = [10 \ 0], J_k = [-10 \ 0] \quad (9.4.26)$$

and, for $\Delta t = .1$,

$$v_k = 10 \int_{t_{k-1}}^{t_k+1} v(t) dt \quad (9.4.27)$$

Recall that

$$E[v(t)v(\tau)] = R\delta(t - \tau) = .02\delta(t - \tau) \quad (9.4.28)$$

We can now calculate the corresponding R_k in the discrete model.

$$\begin{aligned} R_k &= E[v_k^2] = (10)^2 \iint_{\Delta t \text{ interval}} E[v(t)v(\tau)] dt d\tau \\ &= (10)^2(0.02)(.1) = .2 \end{aligned} \quad (9.4.29)$$

The discrete delayed-state model is now complete.

It is somewhat laborious to carry the delayed-state solution through a large number of steps numerically. We can, however, proceed through one step with only modest effort, and this will provide a limited comparison with the other models being considered. Thus consider the initial step from 0 to .1 sec. The first preaveraged measurement is not available until $t = .1$, and so we simply project the initial error covariance matrix ahead in accordance with

$$\mathbf{P}_1^- = \phi_k \mathbf{P}_0 \phi_k^T + \mathbf{Q}_k \quad (9.4.30)$$

where ϕ_k and \mathbf{Q}_k are given by Eqs. (9.4.23) and (9.4.24), and the initial \mathbf{P}_0 is

$$\mathbf{P}_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad (9.4.31)$$

(Note that x_1 is zero at $t = 0$ because of the definition of x_1 , and it is known perfectly.) Numerical evaluation of Eq. (9.4.30) yields

$$\mathbf{P}_1^- \approx \begin{bmatrix} .0096748 & .095163 \\ .095163 & 1.0 \end{bmatrix} \quad (9.4.32)$$

It is now a routine matter to substitute into Eqs. (9.4.11) and (9.4.12) to find the gain and updated \mathbf{P}_1 matrix. The results are

$$\begin{aligned} \mathbf{K}_1 &\approx \begin{bmatrix} .096748 \\ .95163 \end{bmatrix} / (.96748 + .2) \\ &\approx \begin{bmatrix} .082869 \\ .815115 \end{bmatrix} \end{aligned} \quad (9.4.33)$$

and

$$\mathbf{P}_1 \approx \begin{bmatrix} - & - \\ - & .22432 \end{bmatrix} \quad (9.4.34)$$

Only the ‘‘22’’ term of the \mathbf{P}_1 matrix is shown, because it is the only variance of interest. Note that it is only slightly greater than the optimal value of .22069 obtained with the continuous filter.

Usual Kalman Filter with Only One State Variable

We will now treat the averaged measurements just as if they were truly instantaneous measurements occurring at the sample points $t = .1, .2, \dots$, etc. (See the sample function shown in Fig. 5.10 to get some idea as to the grossness of this approximation.) The additive noise variance R_k is .2 just as in the delayed-state model. Even though we know the measurement model is not exactly correct, it is still of interest to compute the resulting gain and associated P at $t = .1$. The process model parameters for this case are

$$\begin{aligned} \phi_k &= e^{-1} \approx .9048374 \\ Q_k &= 1 - e^{-2} \approx .1812692 \end{aligned}$$

The a priori P at $t = .1$ is then

$$\mathbf{P}_1^- = \phi_k \mathbf{P}_0 \phi_k^T + Q_k = 1.0 \quad (9.4.35)$$

(which was more or less obvious at the outset). The gain and a posteriori P_1 are obtained from the usual Kalman filter equations with H_k set equal to 1. Their numerical values work out to be

$$K_1 = \frac{\mathbf{P}_1^-}{\mathbf{P}_1^- + R_1} = \frac{1.0}{1.0 + .2} \approx .83333 \quad (9.4.36)$$

$$P_1 = (1 - KH)P^- \approx .16667 \quad (9.4.37)$$

It is of interest to note that the gain of .83333 is not radically different from the correct gain of .815115 found in the delayed-state model. However, the computed P_1 of .16667 differs considerably from the optimal variance of .22432 computed previously. Thus, the oversimplified one-state model yields an unduly optimistic figure of merit for the estimation error.

Note that in this case we have intentionally refrained from referring to P as the error variance. It is not! It is, in fact, an erroneous figure of merit because of the rather gross approximation made in the measurement model. However, the fact that the gain worked out to be about ‘‘right’’ makes one think that the filter (with its admittedly erroneous model) might still be a reasonably good estimator. The delayed-state model provides a convenient means of assessing the degree of suboptimality in this case. All we need to do is let \mathbf{K}_k in Eq. (9.4.8) be

$$\mathbf{K}_1 = \begin{bmatrix} 0 \\ .83333 \end{bmatrix} \quad (9.4.38)$$

This is the effective suboptimal gain produced by the one-state model. Equation (9.4.8) then yields the actual covariance realized by using the suboptimal

gain in the "truth model." The result is

$$\mathbf{P}_1 \text{ (actual)} = \begin{bmatrix} - & - \\ - & .22470 \end{bmatrix} \quad (9.4.39)$$

Surprisingly, the resultant value of .22470 is only slightly larger than the optimal value of .22432 (which, in turn, was only slightly larger than .22070 for the continuous filter).

There are two main results of this brief analysis:

1. The P_k computation of the simplified one-state model yields an overly optimistic estimate of the filter's accuracy.
2. However, in spite of the modeling errors, the one-state filter's performance is nearly as good as the optimal two-state filter which uses the delayed-state measurement model. Roughly speaking, the P computation is bad, but the estimate is still fairly good.

The preceding conclusions are based on the results of just one recursive step. In order to be confident of these results, we should cycle through the recursive equations until the steady-state situation is reached. This is routine, but it involves considerable effort, so we will leave the example at this point. ■

It should be clear from this example that the delayed-state model is useful as an analysis tool in assessing the degree of suboptimality caused by compressing the measurement data, even if the delayed-state algorithm is not actually implemented on-line. Without this kind of suboptimality analysis, the effect of preaveraging the continuous measurement data has to be guesswork. With the analysis, one can compute quantitative measures of the degree of suboptimality introduced by averaging over any specified interval.

9.5 STOCHASTIC LINEAR REGULATOR PROBLEM AND THE SEPARATION THEOREM

The linear regulator problem is a classical problem of optimal control theory, and it can be posed (and solved) in either continuous or discrete form (13 to 17). In the interest of brevity we will confine our remarks to the continuous version of the problem.

The *deterministic* linear regulator problem may be stated as follows: Given a linear dynamical system of the form

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{B}\mathbf{u} \quad (9.5.1)$$

what $\mathbf{u}(t)$ will minimize the quadratic performance index

$$J = \mathbf{x}^T(t_f) \mathbf{S} \mathbf{x}(t_f) + \int_{t_0}^{t_f} [\mathbf{x}^T(t) \mathbf{W}_x \mathbf{x}(t) + \mathbf{u}^T(t) \mathbf{W}_u \mathbf{u}(t)] dt \quad (9.5.2)$$

where \mathbf{S} , \mathbf{W}_x , and \mathbf{W}_u are symmetric positive definite weighting matrices that are chosen to fit the situation at hand? The intent of the regulator is to reduce the initial state of the system to zero (or nearly so) quickly and without undue control effort. The weighting factor \mathbf{S} penalizes the system for not reaching zero at the specified terminal time t_f . The weighting factors \mathbf{W}_x and \mathbf{W}_u apply penalties for the trajectories of the state $\mathbf{x}(t)$ and the control $\mathbf{u}(t)$ over the time span $[t_0, t_f]$. It can be seen that if \mathbf{W}_x is large and \mathbf{W}_u small, the optimal system will apply a large control effort and force the system toward zero rapidly in order to avoid a large penalty due to large \mathbf{W}_x . On the other hand, if \mathbf{W}_x is small and \mathbf{W}_u large, the system will be frugal with its control effort, and it will approach zero slowly. Clearly, a wide variety of situations can be accommodated within the structure of this formulation. We will not derive the solution of the linear regulator problem here. This is adequately covered in the mentioned references (and many others, as well). We simply state that the optimal control law is a linear feedback law that specifies $\mathbf{u}(t)$ to be

$$\mathbf{u}(t) = -\mathbf{K}_1(t)\mathbf{x}(t) \quad (9.5.3)$$

where the feedback gain $\mathbf{K}_1(t)$ is computed from the system parameters. (It is not a function of \mathbf{x} .) We need not be concerned with the detailed solution for $\mathbf{K}_1(t)$ (except to note parenthetically that there is a close duality between this and the optimal estimator problem.) It is important to note, though, that it is assumed that the state vector \mathbf{x} is available for feedback purposes as indicated in Fig. 9.4. However, in many physical situations, we do not have the privilege of observing all the elements of \mathbf{x} directly; quite to the contrary, we are usually only allowed to observe \mathbf{x} through some output relationship, say

$$\mathbf{y} = \mathbf{H}\mathbf{x} \quad (9.5.4)$$

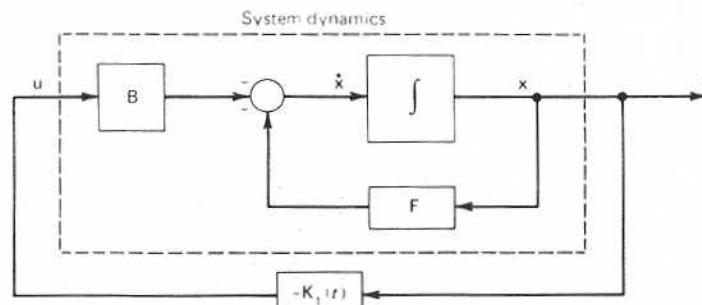


Figure 9.4. Optimal linear regulator.

Now, based on y , we must somehow reconstruct x . If the observation is essentially error free, we can use observer theory in the reconstruction (17 to 20). However, if the observation of x is corrupted with noise, as is often the case, then the reconstruction of x becomes an estimation problem. This leads to the stochastic linear regulator problem,* which will now be formulated.

Let the system dynamics be specified by the linear equation

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{G}\mathbf{w} \quad (9.5.5)$$

where \mathbf{F} , \mathbf{B} , and \mathbf{u} are as before, and \mathbf{w} is an additional Gaussian white noise forcing function that is characterized by

$$E[\mathbf{w}(t)\mathbf{w}^T(\tau)] = \mathbf{O}\delta(t - \tau) \quad (9.5.6)$$

The system state vector is assumed to be observed via the relationship

$$\mathbf{z} = \mathbf{Hx} + \mathbf{v} \quad (9.5.7)$$

where \mathbf{v} is Gaussian white noise and is characterized by

$$E[\mathbf{v}(t)\mathbf{v}^T(\tau)] = \mathbf{R}\delta(t - \tau) \quad (9.5.8)$$

The two white noises w and v will be assumed to be independent, and, in order to avoid any questions about singular conditions, we assume the system is controllable with respect to the control input u and is observable with respect to z . The optimization problem is to minimize the following cost function:

$$J = E\{\mathbf{x}^T(t_f)\mathbf{S}\mathbf{x}(t_f) + \int_{t_i}^{t_f} [\mathbf{x}^T(t)\mathbf{W}_x\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{W}_u\mathbf{u}(t)] dt\} \quad (9.5.9)$$

As before, we will not dwell on the details of the solution; we will simply state the results (15,16,17,19). Design of the optimal stochastic linear regulator may be separated into two steps:

- First, design the minimum mean-square error estimator for \mathbf{x} , treating \mathbf{u} just as if it were a known deterministic input. The optimal estimator is, of course, a Kalman filter with parameters \mathbf{F} , \mathbf{GQG}^T , \mathbf{H} , and \mathbf{R} . Note that the process has a known input as well as a random input; therefore, the differential equation for $\hat{\mathbf{x}}$ is

$$\dot{\hat{x}} = F\hat{x} + Bu + K(z - H\hat{x}) \quad (9.5.10)$$

where \mathbf{K} is the continuous Kalman gain (see Problem 7.8).

2. Next, solve the deterministic linear regulator problem for the optimal feedback gain $K_1(t)$ just as if a perfect measurement of x were available and $w(t)$ were absent. Then let control input u be

$$\mathbf{u} = -\mathbf{K}_1(t)\hat{\mathbf{x}}(t) \quad (9.5.11)$$

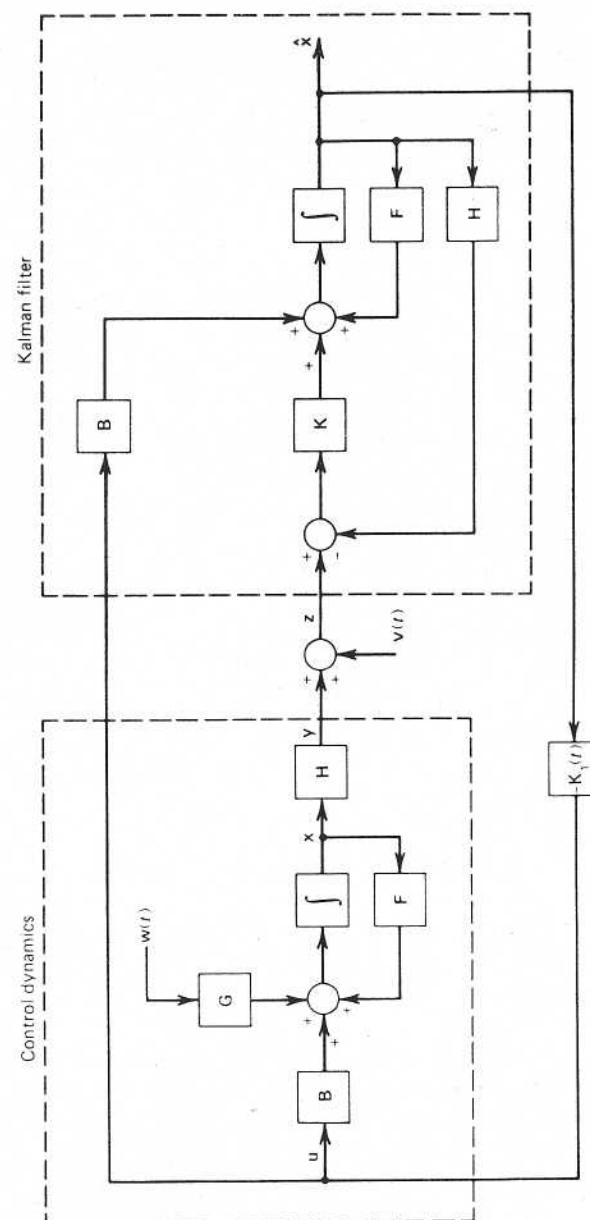


Figure 9.5. Optimal stochastic linear regulator.

*This problem is also called the linear quadratic gaussian (LOG) problem.

and the cost function is minimized. The resulting optimal system is summarized in Fig. 9.5.

The two-step solution just described is known as the *separation theorem* or *separation principle*. It is a most remarkable result in that we would normally expect that the feedback loop would horribly complicate matters by mixing together the control and estimation problems. However, it does not; the two problems separate nicely! The superficial reason for this is the duality between the optimal control and optimal filter problems. The underlying reason for the duality in the first place, though, is not all that obvious, so we will simply say that this is a fortuitous circumstance.

It is not our intent here to teach the subject of optimal control. The subject is well developed, and many fine books have been written in this area in recent years. We simply want to point out with this limited example that estimation theory (and Kalman filtering in particular) plays an important role in optimal control theory, and it behooves the control systems engineer/scientist to understand the rudiments of estimation theory.

Problems

9.1. In Example 9.1, the linearized \mathbf{F} matrix was derived using incremental methods. Show that the same matrix may be obtained from the $\partial\mathbf{f}/\partial\mathbf{x}$ formula [Eq. (9.1.8)]. Note that in Eq. (9.1.8) you must let the *total* positions and velocities, that is, r , \dot{r} , θ , and $\dot{\theta}$, be the state variables.

9.2. Consider a simple one-dimensional trajectory determination problem as follows. A small object is launched vertically in the earth's atmosphere. The initial thrust exists for a very short time, and the object "free falls" ballistically for essentially all of its straight-up, straight-down trajectory. Let y be measured in the up direction, and assume that the nominal trajectory is governed by the following dynamical equation:

$$m\ddot{y} = -mg - D\dot{y}|\dot{y}|$$

where

$$m = .05 \text{ kg} \quad (\text{mass of object})$$

$$g = 9.807 \text{ m/sec}^2 \quad (\text{acceleration of gravity})$$

$$D = 1.4 \times 10^{-4} \text{ N/(m/sec)}^2 \quad (\text{drag coefficient})$$

The drag coefficient will be assumed to be constant for this relatively short trajectory, and note that drag force is proportional to $(\text{velocity})^2$, which makes the differential equation nonlinear. Let the initial conditions be as follows:

$$y(0) = 0$$

$$\dot{y}(0) = 85 \text{ m/sec}$$

The body is tracked and noisy position measurements are obtained at intervals of .1 sec. The measurement error variance is $.25 \text{ m}^2$. The actual trajectory will differ from the

nominal one because of uncertainty in the initial velocity and inaccuracies in the drag model. Assume that the initial position is known perfectly but that initial velocity is best modeled as a normal random variable described by $N(85 \text{ m/sec}, 1 \text{ m}^2/\text{sec}^2)$. Work out the linearized discrete Kalman filter model for the up portion of the trajectory. (*Hint:* An analytical solution for the nominal trajectory may be obtained by considering the differential equation as a first-order equation in velocity. Note $|\dot{y}| = \dot{y}$ during the up portion of the trajectory. Since variables are separable in the velocity equation, it can be integrated. The velocity can then be integrated to obtain position.)

- 9.3.** (a) At the k th step of the usual nonadaptive Kalman filter, the measurement residual is $(z_k - \mathbf{H}_k \hat{\mathbf{x}}_k)$. Let z_k be scalar and show that the expectation of the squared residual is minimized if $\hat{\mathbf{x}}_k^-$ is the optimal a priori estimate of \mathbf{x}_k , that is, the one normally computed in the projection step of the Kalman filter loop. (*Hint:* Use the measurement relationship $z_k = \mathbf{H}_k \mathbf{x}_k + v_k$ and note that v_k and the a priori estimation error are uncorrelated. Also note that the a priori estimate $\hat{\mathbf{x}}_k^-$, optimal or otherwise, can only depend on the measurement sequence up through z_{k-1} and not z_k .)
- (b) Show that the time sequence of residuals $(z_k - \mathbf{H}_k \hat{\mathbf{x}}_k), (z_{k+1} - \mathbf{H}_{k+1} \hat{\mathbf{x}}_{k+1}), \dots$ is a white sequence if the filter is optimal. (*Hint:* See Gelb (21), Chap. 9, for help. Gelb also has an interesting adaptive filter example that makes use of the white sequence property.)

References Cited in Chapter 9

1. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Trans. of the ASME, Jour. of Basic Engr.*, 35-45 (March 1960).
2. R. E. Kalman, and R. S. Bucy, "New Results in Linear Filtering and Prediction," *Trans. of the ASME, Jour. of Basic Engr.*, 83: 95-108 (1961).
3. T. Kailath, "A View of Three Decades of Linear Filtering Theory," *IEEE Trans. on Information Theory*, IT-20: No. 2, 146-181 (March 1974).
4. H. W. Sorenson, "Kalman Filtering Techniques," in *Advances in Control Systems*, Vol. 3, C. T. Leondes (Ed.), New York: Academic Press, 1966, pp. 219-289.
5. R. G. Brown, and J. W. Nilsson, *Introduction to Linear Systems Analysis*, New York: Wiley, 1962.
6. D. T. Magill, "Optimal Adaptive Estimation of Sampled Stochastic Processes," *IEEE Trans. on Automatic Control*, AC-10: No. 4, 434-439 (October 1965).
7. G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, New York: Academic Press, 1977.
8. P. S. Maybeck, *Stochastic Models, Estimation and Control*, Vol. 1, New York: Academic Press, 1979.
9. R. H. Battin, *Astronautical Guidance*, New York: McGraw-Hill, 1964, pp. 338-340.
10. R. G. Brown, and G. L. Hartman, "Kalman Filter with Delayed States as Observables," *Proc. of the National Electronics Conference*, Chicago, Ill., 1968.
11. R. G. Brown, "Analysis of an Integrated Inertial/Doppler-Satellite Navigation System: Part I, Theory and Mathematical Model," Tech. Report No. ERI 62600, Engr. Research Inst., Iowa State University, Ames, 1969.

12. R. G. Brown, and L. L. Hagerman, "An Optimum Inertial Doppler-Satellite Navigation System," *Navigation, The Jour. of the Inst. of Navigation*, 16: No. 3, 260-269 (Fall 1969).
13. M. Athans, and P. L. Falb, *Optimal Control*, New York: McGraw-Hill, 1966.
14. D. E. Kirk, *Optimal Control Theory*, Englewood Cliffs, N.J.: Prentice-Hall, 1970.
15. J. S. Meditch, *Stochastic Optimal Linear Estimation and Control*, New York: McGraw-Hill, 1969.
16. A. E. Bryson, Jr., and Y. Ho, *Applied Optimal Control*, Rev. ed., New York: Halsted Press, Div. of John Wiley & Sons, 1975.
17. A. P. Sage, and C. C. White, *Optimum Systems Control*, 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1977.
18. C. T. Chen, *Introduction to Linear System Theory*, New York: Holt, Rinehart and Winston, 1970.
19. G. F. Franklin, and J. D. Powell, *Digital Control of Dynamic Systems*, Reading, Mass.: Addison-Wesley, 1980.
20. T. Kailath, *Linear Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1980.
21. A. Gelb (Ed.), *Applied Optimal Estimation*, Cambridge, Mass.: MIT Press, 1974.

APPENDIX

Laplace and Fourier Transforms

Elementary treatments of Laplace and Fourier transforms usually gloss over matters of convergence and formal inversion of these transforms by the inversion integral. There are problems in signal analysis where these matters are important though (e.g., Wiener filter theory), and thus we will embellish on these ideas here. It is not the intent here to teach linear transform theory from the beginning. We assume that the reader has the usual manipulative skills in Laplace and Fourier transforms that one would normally get in an undergraduate electrical engineering program. [For example, see Shanmugam (1) or Carlson (2) for Fourier transforms and Dorf (3) or D'Azzo and Houpis (4) for Laplace transforms.] Here the emphasis will be to place Laplace and Fourier transforms in perspective relative to each other, and to discuss, in particular, formal inversion of these transforms by the inversion integral. We begin with the one-sided Laplace transform.

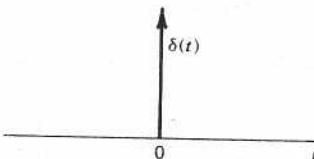
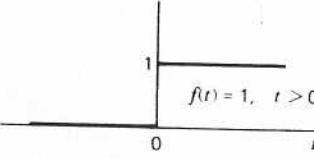
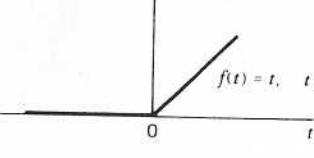
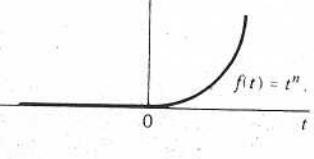
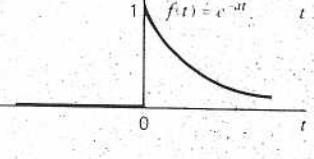
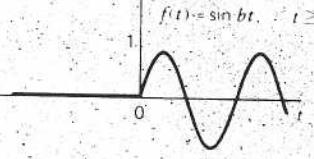
A.1 THE ONE-SIDED LAPLACE TRANSFORM

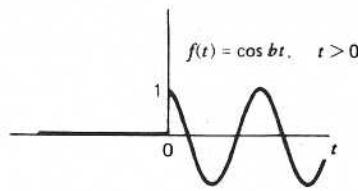
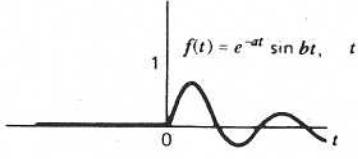
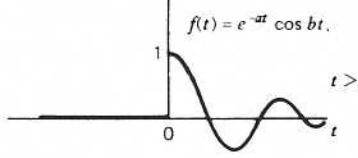
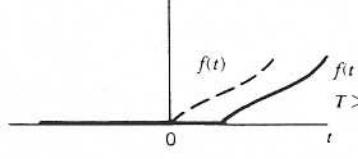
Electrical engineers usually first encounter Laplace transforms in circuit analysis, and then again in linear control theory. In both cases the central problem is one of finding the system response to an input initiated at $t = 0$. Since the time history of the system prior to $t = 0$ is summarized in the form of the initial conditions, the ordinary one-sided Laplace transform serves us quite well. Recall that it is defined as

$$F(s) = \int_{0+}^{\infty} f(t)e^{-st} dt \quad (\text{A.1})$$

The defining integral is, of course, insensitive to $f(t)$ for negative t : but, for reasons that will become apparent shortly, we arbitrarily set $f(t) = 0$ for $t < 0$ in one-sided transform theory. The integral of Eq. (A.1) has powerful convergence properties because of the e^{-st} term. We know it will always converge somewhere in the right-half s plane, provided that we only consider inputs (and responses) that increase no faster than at some fixed exponential rate. This is usually the case in circuits and control problems, and hence the actual region of convergence is of little concern. A common region of convergence is tacitly

Table A.1 Common One-Sided Laplace Transform Pairs^a

Name	Pictorial Description	Laplace Transform
Unit impulse (Area is to right of origin)		1
Unit step		$\frac{1}{s}$
Unit ramp		$\frac{1}{s^2}$
General integer power of t		$\frac{n!}{s^{n+1}}$
Damped exponential		$\frac{1}{s+a}$
Sine wave		$\frac{b}{s^2 + b^2}$

Name	Pictorial Description	Laplace Transform
Cosine wave		$\frac{s}{s^2 + b^2}$
Damped sine wave		$\frac{b}{(s+a)^2 + b^2}$
Damped cosine wave		$\frac{s+a}{(s+a)^2 + b^2}$
Delayed positive-time function		$F(s)e^{-Ts}$

^a Time functions having a discontinuity at $t = 0$ are intentionally left undefined at the origin in this table. The missing value does not affect the direct transform. See Exercise A.1 at the end of this Appendix for a discussion of the appropriate choice of $f(0)$ to assure compatibility with the inversion integral.

assumed to exist somewhere, and we simply adopt a table look-up viewpoint for getting back and forth between the time and complex s domains. For reference purposes a brief list of common transform pairs is given in Table A.1. Note again that we have intentionally defined all time functions in the table to be zero for $t < 0$. We will have occasion later to refer to such functions as *positive-time* type functions. It is also worth mentioning that the impulse function of one-sided transform theory is considered to have all its area to the right of the origin in the limiting process. Thus it is a positive-time function. (The word function is abused a bit in describing an impulse, but this is common usage, so it will be continued here.)

A.2 THE FOURIER TRANSFORM

The Fourier transform is used widely in communications theory where we often wish to consider signals that are nontrivial for both positive and negative time. Thus a two-sided transform is appropriate. Recall that the Fourier transform of $f(t)$ is defined as

$$F(j\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt \quad (\text{A.2})$$

We know, through the evolution of the Fourier transform from the Fourier series, that $F(j\omega)$ has the physical significance of signal spectrum. The parameter ω in Eq. (A.2) is $(2\pi) \times$ (Frequency in hertz), and in elementary signal analysis we usually consider ω to be real. This leads to obvious convergence problems with the defining integral, Eq. (A.2), and is usually circumvented simply by restricting the class of time functions being considered to those for which convergence exists for real ω . The two exceptions to this are constant (*d-c*) and harmonic (sinusoidal) signals. These are usually admitted by going through a limiting process that leads to Dirac delta functions in the ω domain. Even though the class of time functions allowed is somewhat restrictive, the Fourier transform is still very useful because many physical signals just happen to fit into this class (e.g., pulses and finite-energy signals). If we take convergence for granted, we can form a table of transform pairs, just as we did with Laplace transforms, and Table A.2 gives a brief list of common Fourier transform pairs.

For those who are more accustomed to one-sided Laplace transforms than Fourier transforms, there are formulas for getting from one to the other. These are especially useful when the time functions have either even or odd symmetry. Let $f(t)$ be a time function for which the Fourier transform exists, and let

$\mathcal{F}[f(t)]$ = Fourier transform of $f(t)$

$F(s)$ = One-sided Laplace transform of $f(t)$

Then, if $f(t)$ is even,

$$\mathcal{F}[f(t)] = F(s)|_{s=j\omega} + F(s)|_{s=-j\omega} \quad (\text{A.3})$$

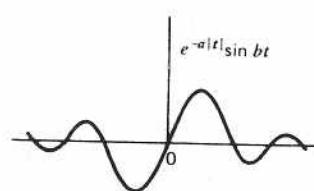
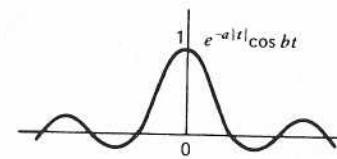
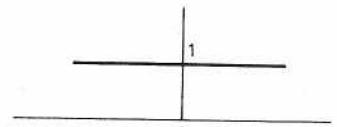
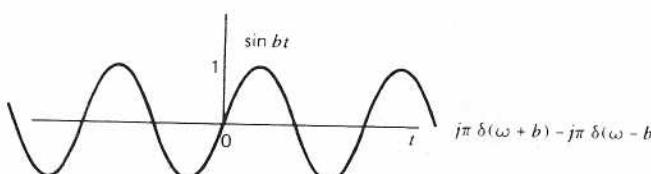
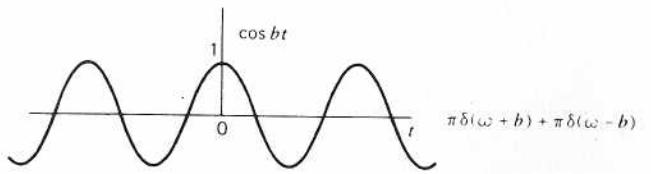
Or, if $f(t)$ is odd,

$$\mathcal{F}[f(t)] = F(s)|_{s=j\omega} - F(s)|_{s=-j\omega} \quad (\text{A.4})$$

These formulas follow directly from the defining integrals of the two transforms.

Table A.2 Common Fourier Transform Pairs

Name	Pictorial Description	Fourier Transform
Damped exponential		$\frac{2a}{\omega^2 + a^2}$
Rectangular pulse		$T \frac{\sin(\omega T/2)}{(\omega T/2)}$
Triangular pulse		$\frac{T}{2} \left[\frac{\sin(\omega T/4)}{\omega T/4} \right]^2$
Gaussian pulse		$\frac{\sqrt{\pi}}{a} e^{-\omega^2/4a^2}$
Symmetric impulse		$\delta(t)$
Sinc function (sinc 2Wt)		$F(j\omega) = \begin{cases} \frac{1}{2W} & \omega \leq 2\pi W \\ 0 & \omega > 2\pi W \end{cases}$

Name	Pictorial Description	Fourier Transform
Damped sine wave		$\frac{ja}{a^2 + (\omega + b)^2}$ $-\frac{ja}{a^2 + (\omega - b)^2}$
Damped cosine wave		$\frac{a}{a^2 + (\omega + b)^2}$ $+\frac{a}{a^2 + (\omega - b)^2}$
Constant		$2\pi \delta(\omega)$
Sine wave		$j\pi \delta(\omega + b) - j\pi \delta(\omega - b)$
Cosine wave		$\pi \delta(\omega + b) + \pi \delta(\omega - b)$

A.3 TWO-SIDED LAPLACE TRANSFORM

The usual Fourier transform with ω real serves us well for much of signal analysis, but there are occasions where we wish to consider functions for which this transform does not exist. In such cases the so-called two-sided Laplace transform is sometimes useful. It is defined as

$$F(s) = \int_{-\infty}^{\infty} f(t)e^{-st} dt \quad (A.5)$$

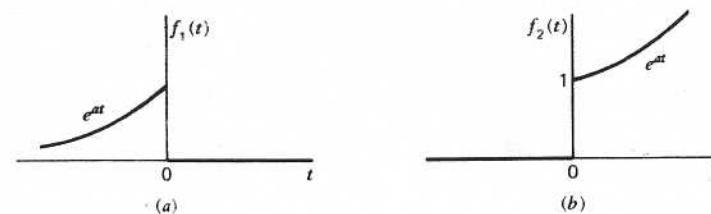


Figure A.1. Time functions for Example A.1. (a) Negative-time function. (b) Positive-time function.

It can be seen to be identical in form to the Fourier transform except for notation, that is, $j\omega$ is replaced with s . Innocent as this change may appear at first glance, it is important, because we now wish to let s be complex. Furthermore, we do not wish to place any restrictions on the type of time functions being considered, other than to say the integral of Eq. (A.5) must converge *somewhere* in the s plane. Now the region of convergence becomes important, and an example will illustrate this.

Example A.1 Consider the two time functions shown in Fig. A.1. Their respective two-sided Laplace transforms may be found from Eq. (A.5).

For signal $f_1(t)$ of Fig. A.1a,

$$F_1(s) = \int_{-\infty}^0 e^{at} e^{-st} dt = \frac{-1}{s - a}, \quad \text{for } \operatorname{Re}[s] < a$$

For signal $f_2(t)$ of Fig. A.1b,

$$F_2(s) = \int_0^{\infty} e^{at} e^{-st} dt = \frac{1}{s - a}, \quad \text{for } \operatorname{Re}[s] > a$$

Here we have the uncomfortable situation of two different time functions having the same transform except for sign! There is a saving feature though—their regions of convergence are nonoverlapping. Hence, if we add the qualification of the convergence region to the functional form in the s domain, we then restore the one-to-one transform-pair relationship that we must have in any transform algebra. In this case the entry in our table of transform pairs might appear as follows:

Time Function	Transform
$f(t) = \begin{cases} e^{at}, & t < 0 \\ 0, & t > 0 \end{cases}$	$\frac{-1}{s - a}$ and region of convergence is $\operatorname{Re}[s] < a$
$f(t) = \begin{cases} 0, & t < 0 \\ e^{at}, & t > 0 \end{cases}$	$\frac{1}{s - a}$ and region of convergence is $\operatorname{Re}[s] > a$

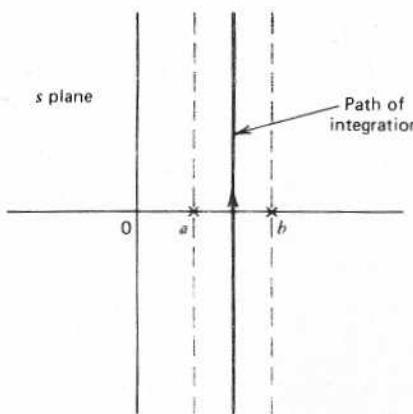


Figure A.2. Convergence strip and path of integration for Example A.2.

Tacking on the region of convergence in our table of transform pairs is a bit cumbersome, but it must be done, or at least tacitly implied somehow.

Table A.3 shows a few common time functions, their two-sided Laplace transforms, and their respective regions of convergence. The list is by no means complete. It is simply intended to show the variety of situations that can occur. Note that a number of the entries have regions of convergence that do not include the $j\omega$ axis, and therefore usual Fourier transforms will not exist for these time functions. Also, two of the time functions shown do not have two-sided Laplace transforms, because the defining integral does not converge anywhere in the s plane. ■

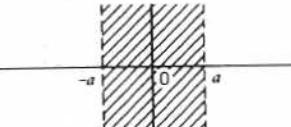
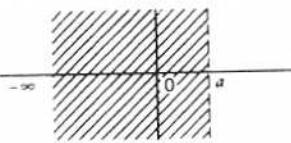
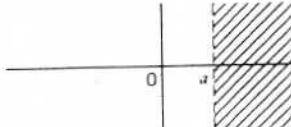
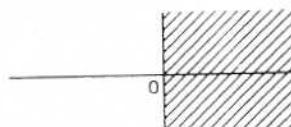
The Inversion Integral

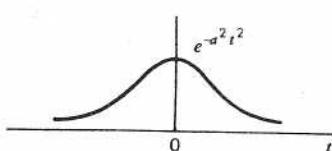
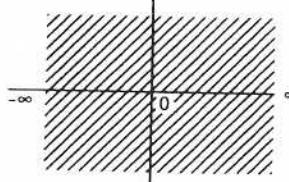
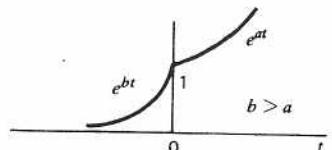
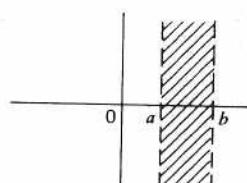
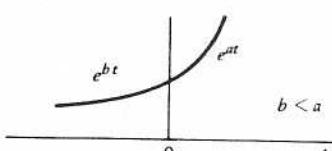
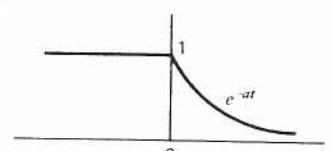
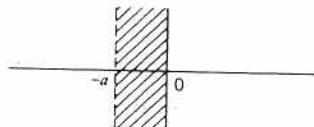
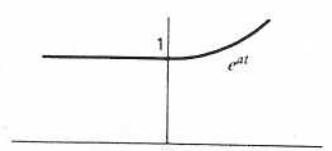
To compile and work with two-sided transform tables such as Table A.3 would be awkward, to say the least. The addition of the specification of convergence region to the functional form is like adding an extra dimension to the transform. Thus, rather than use cumbersome two-sided transform tables for inversion, it is more convenient to use one-sided transform tables with the appropriate interpretation as to positive- and negative-time parts of the time function. This is where the formal rules of inversion are helpful. Beginning with the Fourier transform, we have the inversion formula

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega \quad (\text{A.6})$$

Recall that this came from a limiting situation of the complex Fourier series where the expansion interval was allowed to approach infinity. The corre-

Table A.3 Two-Sided Laplace Transforms

Time Function	Two-sided Laplace Transform	Region of Convergence
e^{-at}	$\frac{1}{s+a}$	
$t^n e^{-at}$	$\frac{n!}{(s+a)^{n+1}}$	
t^n	$\frac{n!}{s^{n+1}}$	
e^{at}	$\frac{1}{s-a}$	
$\delta(t)$	1	
$\delta(t-a)$	e^{-as}	

Time Function	Two-sided Laplace Transform	Region of Convergence
	$\frac{\sqrt{\pi}}{a} e^{(s^2/4a^2)}$	
	$\frac{1}{s-a} + \frac{1}{-s+b}$	
	---	Converges nowhere in s plane
	$\frac{1}{s+a} + \frac{1}{-s}$	
	---	Converges nowhere in s plane

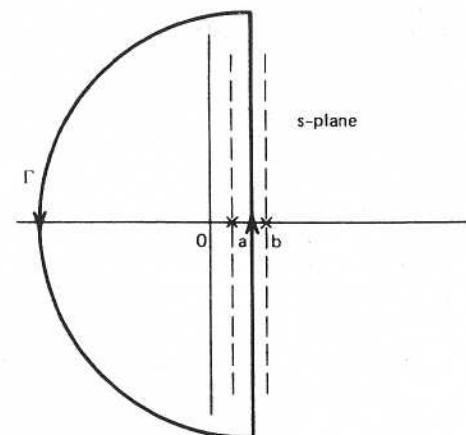


Figure A.3. Closing the path to left to get the positive-time part of $f(t)$.

sponding inversion integral for the Laplace transform is

$$f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s)e^{st} ds \quad (\text{A.6})$$

This formula applies to both the one- and two-sided transforms. Note that the integration is along a vertical line in the s plane, and the constant c is chosen such that *the path of integration lies within the strip of convergence of $F(s)$* . Knowledge of the convergence region is the key to getting the correct inverse transform (i.e., time function) in the inversion process. An example will illustrate this.

Example A.2 Consider the transform pair shown as the seventh entry in Table A.3. The transform and the strip of convergence are

$$F(s) = \frac{1}{s-a} + \frac{1}{-s+b}; \quad a < \text{Re}[s] < b$$

The theory says to choose the path of integration between the poles at a and b as shown in Fig. A.2. One could evaluate the integral by treating the complex function as a sum of real and imaginary parts, and then use real variable calculus for each part. This, however, is working the problem the hard way (see Exercise A.1). An easier way is by means of residue theory (5). It can be shown (6) that the addition of an infinite semicircular path to the left as shown in Fig. A.3 contributes nothing to the integral of Eq. (A.6) for positive t . (This can be seen intuitively by noting that the real part of s has a large negative value along this path.) Thus the positive-time part may be evaluated by the method of residues as follows:

$$\begin{aligned}
 f(t) &= \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s) e^{st} dt \\
 &= \frac{1}{2\pi j} \cdot 2\pi j [\text{Sum of residues of } F(s) e^{st} \text{ within contour } \Gamma] \\
 &= \left[\left(\frac{1}{s-a} + \frac{1}{-s+b} \right) (e^{st}) (s-a) \right]_{s=a} \\
 &= e^{at} \quad (\text{for } t > 0 \text{ only})
 \end{aligned}$$

Similarly, it should be apparent that if we close the contour to the right, we get the negative-time portion of $f(t)$. In this example it is

$$\begin{aligned}
 f(t) &= -\frac{1}{2\pi j} \cdot 2\pi j [\text{Sum of residues of } F(s) e^{st} \text{ within} \\
 &\quad \text{clockwise contour closed to the right}] \\
 &= \left[-\left(\frac{1}{s-a} + \frac{1}{-s+b} \right) (e^{st}) (s-b) \right]_{s=b} \\
 &= e^{bt} \quad (\text{for } t < 0 \text{ only}) \quad \blacksquare
 \end{aligned}$$

The procedure illustrated in Example A.2 can be simplified even further by simply noting that poles of $F(s)$ to the left of the vertical path of integration contribute to the positive-time part of $f(t)$, and those to the right yield the negative-time part. Thus, if $F(s)$ is rational, a simpler inversion algorithm may be stated:

1. First think of writing $F(s)$ in terms of a partial fraction expansion and grouping terms together in the form

$$F(s) = \left[\begin{array}{l} \text{Terms with poles to left} \\ \text{of path of integration} \end{array} \right] + \left[\begin{array}{l} \text{Terms with poles to right} \\ \text{of path of integration} \end{array} \right] \quad (\text{A.7})$$

Note that $F(s)$ is to be resolved into an *additive* combination, and thus this is not the same as spectral factorization.

2. Then the inverse transform of $F(s)$ is obtained by evaluating the positive- and negative-time parts separately. The first term of Eq. (A.7) yields the positive-time part, and the second gives the negative-time part. Each may be evaluated using one-sided transform methods. (See Exercise A.1 for a discussion of the evaluation of $f(t)$ at $t = 0$.)

Note that if there are no poles to the right of the path of integration (e.g., one-sided Laplace transforms), the inversion integral automatically yields zero for the negative-time part. This is why we must consider $f(t)$ to be zero for negative time in one-sided transform theory, even though the defining integral is insensitive to the negative-time portion of the function. Also note that if the

strip of convergence includes the imaginary axis, the two-sided Laplace transform degenerates to the Fourier transform with a simple change in notation (i.e., $j\omega$ is replaced with s). Thus resolving the transform into positive- and negative-time parts is also useful in the evaluation of inverse Fourier transforms.

Our presentation of Laplace and Fourier transforms has been largely an overview, and it was intended to place these transforms in proper perspective relative to each other. Many details have thus been omitted. Also, there are other linear transforms that are of interest in signal analysis, and these have not even been mentioned. Thus, much of linear transform theory has been left unsaid here. For further reading, see Bracewell (7). This is an especially readable book on the subject and it includes many applications.

Exercise A.1

It was stated earlier that the inversion integral automatically yields zero for $t < 0$ in one-sided transform theory. This statement becomes more convincing after actually performing the integration for a sample problem. Consider the one-sided transform pair:

$$f(t) = \begin{cases} e^{-2t}, & t > 0 \\ 0, & t < 0 \end{cases} \quad (\text{A.8})$$

$$F(s) = \frac{1}{s+2}, \quad \text{Re}[s] > -2 \quad (\text{A.9})$$

Obviously the "strip" of convergence is the entire semi-infinite s plane to the right of $\text{Re}[s] = -2$. Thus it includes the $j\omega$ axis. A legitimate inversion integral would then be

$$\text{Inverse of } F(s) = \frac{1}{2\pi j} \int_{-\infty}^{\infty} \frac{1}{j\omega + 2} e^{j\omega t} d(j\omega) \quad (\text{A.10})$$

- Rewrite Eq. (A.10) as the sum of real and imaginary parts, and then integrate each part separately using ordinary real calculus methods. The end result should be just $f(t)$ as stated in the transform-pair statement. (*Hint:* Take advantage of even and odd symmetry wherever possible. Also, you may find it necessary to use tables of integrals in evaluating the resulting real integrals. The tables given in Dwight (8) are adequate for this problem.)
- In Eq. (A.8) the value of $f(t)$ was intentionally left undefined at the point of discontinuity, that is, at $t = 0$. Find the value of $f(t)$ at $t = 0$ as dictated by the inversion integral, Eq. (A.10). Does this seem reasonable in terms of Fourier series theory? (*Hint:* Evaluate the Fourier series for a square wave at the point of discontinuity and compare the result with that obtained from the inversion integral.)

References

1. K. S. Shanmugam, *Digital and Analog Communication Systems*, New York: Wiley, 1979.
2. A. B. Carlson, *Communication Systems*, 2nd ed., New York: McGraw-Hill, 1975.
3. R. C. Dorf, *Modern Control Systems*, 3rd ed., Reading, Mass.: Addison-Wesley, 1980.
4. J. J. D'Azzo and C. H. Houpis, *Linear Control System Analysis and Design*, 2nd ed., New York: McGraw-Hill, 1981.
5. R. V. Churchill, J. W. Brown, and R. F. Verhey, *Complex Variables and Applications*, 3rd ed., New York: McGraw-Hill, 1976.
6. S. Goldman, *Transformation Calculus and Electrical Transients*, Englewood Cliffs, N.J.: Prentice-Hall, 1949.
7. R. N. Bracewell, *The Fourier Transform and Its Applications*, 2nd ed., New York: McGraw-Hill, 1978.
8. H. B. Dwight, *Tables of Integrals and Other Mathematical Data*, 4th ed., New York: Macmillan, 1961.

Index

- Adaptive Kalman filter, 302
Aided inertial navigation, 222
Anderson, W. G., 167
Augmenting state vector, 213
Autocorrelation function, 79–83
 definition, 79
 general properties, 82
 time autocorrelation, 81
Average, 22, 24
Ballistic trajectory example, 324
Bandlimited systems, 127
Bandlimited white noise, 91
Baro-inertial altimeter example, 177
Bayes rule, 13, 34, 37
Bierman, G. J., 307, 312
Biomedical instrumentation example, 175
Bivariate normal density, 38, 53
Bivariate random variable, 32
Black, H. S., 109
Bolie, V. W., 176
Bona, B. E., 213
Bracewell, R. N., 339
Brown, R. G., 171, 230
Brownian-motion process, 98
Bryson, A. E., 262
Bucy, R. S., 253
Butterworth filter, 140, 174
Carlson, A. B., 327
Central limit theorem, 43
 discrete example, 65
CEP (circular error probable), 67
Characteristic function, 25, 119
Chen, C. T., 266, 267
Cholesky decomposition, 308
Circular error probable (CEP), 67
Coherence function, 89, 145
Complementary filter, 165, 177, 178
Conditional probability, 13, 34
Continuous Kalman filter, 253–272. *See also*
 Kalman filter, continuous
Convergence of Laplace and Fourier
 transforms, 330–339
Convergence of random variable, 58
 in mean, 61
 in probability, 61
Conversion of discrete measurements, to form,
 312
Coriolis theorem, 238
Correlated measurement and process noise,
 259–262
Correlation coefficient, 40
Covariance matrix, 39, 52. *See also* Error
 covariance matrix
Craps, 3, 63
Crosscorrelation function, 83
Crosscorrelator method of determining filter
 weighting function, 144
Cross spectral density function, 89
Cumulative probability distribution function,
 21, 35
Davenport, W. B., 177
D'Azzo, J. J., 327
Decorrelating measurement errors, 250
Delayed-state Kalman filter, 312–320
Deterministic inputs, 248, 271
Deterministic least squares, 242, 251
Deterministic random process, 79
Differentiation formulas (matrix), 210, 243
Discrete Fourier transform (DFT), 110–114
Discrete Kalman filter, 181–252. *See also*
 Kalman filter, discrete
Discrete time state model, 188
Disjoint events, 6

Dorf, R. C., 327
Dwight, H. B., 339

Equivalent events, 18
Ergodic, 78
Error, circular probable, 67
Error analysis:
continuous Kalman filter, 266
discrete Kalman filter, 239
optimal filter, 239
suboptimal, 239, 266, 312–320

Error covariance matrix:
continuous, 255
discrete, 197, 198
prediction, 274
smoothing, 276

Expected value, 23
Experimental data:
autocorrelation of, 103
spectral density of, 103

Exponential probability density function, 65
Extended Kalman filter, 296, 300

Failure probability, 65
Fast Fourier transform (FFT), 114
Filter problem, 151
Fourier transform, 327–340
discrete, 110
fast, 114
short table, 331
Fraser, D. C., 285, 287
Fried, W. R., 177
Fritze, E. H., 167

Gain:
continuous Kalman, 255
discrete Kalman, 198
smoothing, 276

Gaussian:
narrowband noise, 96, 119
random process, 77
random variable, *see* Normal random variable
white noise, 100

Gauss-Markov process:
definition, 92
simulation examples, 202, 210

Gelb, A., 285, 287, 325
Girgis, A., 230
Global positioning system (GPS), 118, 222
Goren, C. H., 63
Gram-Schmidt orthogonalization procedure, 222

Houpis, C. H., 327

Impulse probability density function, 31, 38
Independence, *see* Independent
Independent:
events, 15
random processes, 76
random variables, 34, 37
sum of, 40

Inertial navigation:
aided systems, 222–229
marine error model, 214, 224
pure systems, 223
single channel error model, 143, 174

Instrument landing system (ILS) example, 167
Integral tables, mean square value, 126
Inversion integral, 334

Jacobian, 49
Johansen, D. E., 262
Joint probability, 10, 11, 33

Kailath, T., 296
Kalman, R. E., 147, 181, 253. *See also*
Kalman filter
Kalman filter:
continuous, 253–272
derivation, 253–257
with deterministic inputs, 271
error covariance, 255
estimator differential equation, 256
example, 258
gain, 255
summary, 256
discrete, 181–252
adaptive, 302
alternative algorithm, 217
conditional density viewpoint, 204
derivation, 194–200
divergence, 233
extended, 296–300
linearized, 296
modeling, 188–195
recursive equations, 200
self-learning, 302
sequential processing of measurements, 220
summary, 200
with deterministic inputs, 248

Kalman gain:
continuous, 255
discrete, 198

Kalman/Wiener filter relationship, 267
Kayton, M., 177

Laplace transform, 327–339
one-sided, 327
table, 328
two-sided, 332
table, 335

Least squares:
deterministic, 242, 251
perspective, 147
relationship to Kalman filter, 242–245

Linearization of nonlinear dynamics, 296
Linearized Kalman filter, 296
Linear quadratic Gaussian (LQG) problem, 322
Linear regulator problem, 320–324
Linear systems:
analysis problem defined, 122
nonstationary analysis, 131, 133
stationary analysis, 123

Long, F. M., 176

Magill, D. T., 302
Marginal probability, 12, 34
Markov process, 92, 159
Mass distribution:
probability, 20, 33
Matrix differentiation formulas, 210, 243
Maybeck, P. S., 307, 312
Mean, 22, 24
Mean square value, of filter output, 126
Measurement equation:
continuous, 253
discrete, 195
Measurement residuals, 305, 325
Meditch, J. S., 275, 280
Minimum-phase transfer function, 130
Modeling errors, effect of, 234
Moment, 23
Multiple-input multiple-output analysis, 137
Multivariate:
random variable, general, 32
normal, 51
Mutually exclusive events, 6

Narrowband Gaussian process, 96, 119
Negative-time function, 156
Nilsson, J. W., 171
Noise equivalent bandwidth, 128
Nondeterministic random process, 79
Normal variable, 26–31
bivariate, 38

central tendency toward, 40
density function, 26
distribution function, 28
linear transformation of, 55
multivariate, 51
tables of density and distribution functions, 28
typical sample values, 211

Nyquist, H., 109
Nyquist rate, 91, 109, 313

Observability, 237, 249
Odds, 62
Off-line error analysis, 239
Ogata, K., 266
Optimization of filter response:
general approach, 151
with respect to parameter, 148

Orthogonality principle, 164, 177
Orthogonal random variables, 39

Periodogram, 85
Phase-lock loop, 145
Phillips, R. S., 126, 148
Phillips optimization procedure, 148, 174
P matrix:
continuous, 255
discrete, 197, 198
Poisson probability distribution, 66, 67
Positive-time function, 156
Potter, J. E., 285, 287, 307
Power relaying application, 229
Power spectral density function, 84–89
Prediction problem:
definition, 151
discrete recursive solution, 273–274
pure prediction example, 158

Probability, 1–70
axiomatic, 5
axioms of, 7
conditional, 13, 34
cumulative distribution function, 21
density function, 21
distribution, 20
distribution function, 21, 33
impulse density function, 31, 38
joint, 10–13, 33
marital, 12, 34
mass distribution, 20, 33
relative frequency viewpoint, 2
space, 8
unconditional, 34

- Pseudorandom:
 binary sequence, 101
 noise, 100
- Ψ equation, 214, 224, 238, 249
- Pure prediction problem, 158
- Q matrix:
 continuous, 253–254
 discrete, 189–195
 numerical evaluation of, 192
- Radio sextant navigation example, 224
- Random constant, estimate of, 177, 270
- Random process, 71–100
 autocorrelation function of, 79–83
 Brownian motion, 98
 deterministic, 79
 Gaussian, 77
 Gauss-Markov, 92
 narrowband, 96, 119
 nondeterministic, 79
 power spectral density of, 84–89
 probabilistic description of, 74–77
 state model of, 183–188
 white, 90
- Wiener, 98
- Random signals:
 introduction, 1
see also Random process
- Random telegraph wave, 94
- Random variable, 17–61
 bivariate, 32
 continuous, 20
 discrete, 20
 Gaussian, 26, 38
 multivariate, 32
 normal, 26, 38
 transformation of, 44, 55
- Random walk, 98, 235
- Rauch, H. E., 275
- Rayleigh probability density function, 50, 67
- Recursive algorithm example, 182
- Relaying application of Kalman filtering, 229
- Reliability, electronic equipment, example, 65
- Residuals:
 deterministic, 242
 measurement, 305, 325
- Riccati equation, 257–259
- Rice, S. O., 1
- R matrix:
 continuous, 253
 discrete, 196
- Root, W. L., 177
- Rough filter, 222

- RTS (Rauch, Tung, Striebel) smoothing algorithm, 276, 292, 294
- Sample mean, 22
- Sample realization of random process, 72
- Sample space, 5
- Sampling theorem, 109
- Satellite example, 298
- Schuler period, 228
- Sequential processing, measurements, 220
- Shanmugam, K. S., 327
- Shannon, C. E., 109
- Shaping filter, 130
- Shift register sequences, 101
- Similarity transformation, 57
- Smay, R. J., 213
- Smoothing, 151, 274–295
 definition, 151
 fixed-interval, 275
 fixed-lag, 283
 fixed-point, 280
 forward-backward filter approach, 285–294
 continuous, 287
 discrete, 288
 initial condition problem, 294
 problem classifications, 275
 Rauch, Tung, Striebel algorithm, 276
- Sorenson, H. W., 197, 245, 298
- Spectral density function, power, 84–89
- Spectral factorization, 125, 156
- Square-root filtering, 307–312
- Standard deviation:
 definition, 24
- State model for:
 Gauss-Markov process, 186
 harmonic motion, 187
 random bias, 187
 random ramp, 187
 stationary process with rational spectral function, 183–186
 Wiener process, 187
- State transition matrix:
 numerical evaluation of, 192
- Stationary process, 77–79, 83
 covariance stationary, 83
 strictly stationary, 78
 wide-sense stationary, 83
- Statistical independence, 34, 37, 76
- Steady state, stationary:
 analysis, 123
- Stochastic linear regulator problem, 320–324
- Stochastic process, 72. *See also* Random process

Suboptimal filter analysis, 240, 266
 Szegö, C. R., 245

- Telegraph wave:
 random, 94
- Telephone load forecasting example, 245
- Thermal noise, 139
- Trace of matrix, 209
 differentiation of, 210
- Transient response, 131, 133
- Truxal, J. G., 174
- Tung, F., 275
- U-D factorization, 311
- Unbiased estimator, 60
- Variance:
 definition, 24
- Vector model of random process, *see* State model
- Venn diagram, 9, 62

- Weighting coefficients for discrete measurements, 171
- White noise, 90
 bandlimited, 91
- Wide-sense stationary, 83
- Wiener filter, 147–180
 multiple-input problem, 169
 nonstationary problem, 160
 perspective, 173
 sampled-data, discrete, problem, 171
 stationary case, 151
 causal solution, 155
 noncausal solution, 153
 two-input problem, 165, 179
- Wiener-Hopf equation, 156
- Wiener/Kalman filter relationship, 267
- Wiener-Khinchine relation, 85
- Wiener, N., 1, 147
- Wiener process, 98
- Wirkler, W. H., 167